Domain Generalization: A Tale of Two ERMs

Anonymous Author(s)
Affiliation
Address
email

Abstract

Domain generalization (DG) is the problem of generalizing from several distributions (or domains), for which labeled training data are available, to a new test domain for which no labeled data is available. A common finding in the DG literature is that it is difficult to outperform empirical risk minimization (ERM) on the pooled training data. In this work, we argue that this finding has primarily been reported for datasets satisfying a *covariate shift* assumption. When the dataset satisfies a *posterior drift* assumption instead, we argue that "domain-informed ERM," wherein feature vectors are augmented with domain-specific information, outperforms pooling ERM. These claims are supported by a theoretical framework and experiments on language and vision tasks.

1 Introduction: the ERM dilemma in domain generalization

Domain generalization (DG) is a learning problem where the learner has access to labeled data from several source domains, and the goal is to generalize to a new target domain for which no labeled data is available. Let X denote the input features, Y the label, and D the domain index.

A persistent puzzle in DG is the surprising effectiveness of empirical risk minimization (ERM), a baseline that simply pools labeled data from all source domains together and trains a domain-agnostic classifier. Despite extensive efforts to design sophisticated DG algorithms, multiple studies have consistently shown that ERM remains highly competitive:

- Gulrajani and Lopez-Paz [2021] (empirical): "when carefully implemented and tuned, ERM outperforms the state-of-the-art in terms of average performance... no algorithm included in DomainBed (dataset) outperforms ERM by more than 1%."
- Rosenfeld et al. [2021] (theory): "IRM and its alternatives fundamentally do not improve over standard Empirical Risk Minimization."
- Teterwak et al. [2025] (empirical): "the additional tuning in our improved baseline ERM++ outperforms both the prior ERM baselines and all recent SOTA methods on DomainBed."

Similar findings about the strong performance of ERM have been reported across other datasets and settings [Koh et al., 2021, Sagawa et al., 2022, Bai et al., 2024].

A related observation is that most existing DG approaches learn a classifier that predicts Y solely from X, thus ignoring domain information during inference. This is reflected in recent surveys:

- Wang et al. [2022] (survey): "The goal of domain generalization is to learn a robust and generalizable predictive function $h : \mathcal{X} \to \mathcal{Y}$ from the M training domains to achieve a minimum prediction error on an unseen test domain S_{test}."
- Zhou et al. [2023] (survey): "The goal of DG is to learn a predictive model $f : \mathcal{X} \to \mathcal{Y}$ using only source domain data such that the prediction error on an unseen domain $T = \left\{x^T\right\}$ is minimized."

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

This is despite the fact that early works on DG learn predictions based not only on feature vectors, but also on domain-specific information [Blanchard et al., 2011, Muandet et al., 2013].

In this work, we argue that conclusions about ERM being "hard to beat" stem primarily from the fact that most benchmark DG datasets are from vision tasks. These datasets are characterized by a *covariate shift* assumption, which means that there is a single classifier that performs well on all domains, and only the marginal distribution of X changes from domain to domain. In such applications, strong performance is indeed possible without the use of domain-specific information, and a domain-agnostic classifier can be trained by ERM on the pooled training data.

Furthermore, we study DG problems characterized by *posterior drift*, where the conditional distribution of $Y|X$ (i.e., the *posterior*) changes with domain. We argue that for such DG problems, pooling ERM is inadequate, and stronger performance is achievable by "domain-informed" ERM, where domain specific information is used both during training and at inference.

The contributions of this work are:

- A theoretical framework extending the original formulation of DG by Blanchard et al. [2011].
- Risk bounds that characterize when domain-specific information is beneficial (posterior drift) and when it is not (covariate shift).
- A quantification of the difference between domain generalization and domain adaptation, addressing an open question in Blanchard et al. [2021, Lemma 9].
- Empirical validation of these findings on both language and vision tasks.

2 Literature review

Blanchard et al. [2011] introduced the domain generalization (DG) problem, motivated by a medical application involving the automatic gating of flow cytometry data. Since then, most DG research has focused on applications in computer vision. A typical DG task in this setting involves training models on labeled images from multiple visual domains (e.g., styles or rendering conditions) and evaluating generalization to a previously unseen domain. Benchmark datasets such as VLCS [Fang et al., 2013], PACS [Li et al., 2017], OfficeHome [Venkateswara et al., 2017], DomainNet [Peng et al., 2019], and ImageNet-Sketch [Wang et al., 2019] have become standard in this line of work.

In these vision-based setups, the underlying distributional shift can be described as covariate shift [Ben-David et al., 2006, Mansour et al., 2009], where the marginal distribution P_X varies significantly across domains—often with disjoint support. Importantly, domain information is frequently viewed as irrelevant or even spurious [Sagawa et al., 2020] for predicting labels. Consequently, much of the literature has focused on learning domain-invariant representations [Sun and Saenko, 2016, Ganin et al., 2016, Arjovsky et al., 2019]. Additional references are in Appendix A.

In contrast, our work is motivated by a different class of problems characterized by *posterior drift* [Scott, 2019, Cai and Wei, 2021, Maity et al., 2024, Zhu et al., 2024], where the conditional distribution of $Y|X$ varies across domains. This type of shift commonly arises in natural language processing (NLP) tasks. For a given input sentence X, different annotators—or populations—may interpret its semantic content differently, leading to divergent labels Y (e.g., offensive vs. non-offensive, positive vs. negative). Such inherent ambiguity in language often results in systematic disagreement in annotations. Empirical studies have documented these effects across a wide range of NLP tasks [De Marneffe et al., 2019, Plank, 2022, Deng et al., 2023].

3 Domain generalization: A general probabilistic formulation

In standard classification, a random pair (X, Y) is assumed to be drawn from a fixed joint distribution P_{XY}, where $X \in \mathcal{X}$ is a feature vector and $Y \in \mathcal{Y} = \{1, \ldots, K\}$ denotes the corresponding class label[1]. The goal is to learn a function $f : \mathcal{X} \to \mathcal{Y}$ that minimizes the risk:

$$\mathbb{E}_{(X,Y)\sim P_{XY}} \left[\mathbb{1}_{\{f(X)\neq Y\}} \right].$$

[1]This section easily extends to regression, but subsequent sections are focused on classification.

Domain generalization (DG) can be framed in a similar way. Let $\mathcal{D}$ denote a set of possible domains, where the term *domain* is a synonym for a joint distribution of X and Y. Let D be a random variable on $\mathcal{D}$. Furthermore, let M be a random variable on a space $\mathcal{M}$ that, intuitively, provides partial information about D. The idea in DG is that D determines a distribution of (X, Y), but is not observed. M provides partial information about D, and is thus useful at test time in adapting the classifier to the test domain. While the choice of M will depend on the application, one choice that is always viable is to take M to be $P_{X|D}$, the marginal distribution X for the given domain, which is known at test time though the unlabeled test sample. As we argue below, the observability of M is what makes DG an interesting problem, and distinct from standard classification.

Formally, we assume that (X, Y, M, D) are jointly distributed, with joint distribution denoted as P_{XYMD}. This distribution induces several other distributions of interest in this paper. We follow convention in denoting marginal distributions by keeping the relevant subscripts. For example, P_{XYD} denotes the joint distribution of (X, Y, D) after M is marginalized out. Similarly, P_{XY} denotes the marginal distribution of (X, Y).

For any fixed $d \in \mathcal{D}$, $P_{XY|D=d}$ is a joint distribution of (X, Y). Note that our notation is somewhat redundant, as both d and $P_{XY|D=d}$ are notations for the same thing – a domain = a joint distribution of (X, Y) – but these two notations will serve different purposes in our discussion.[2]

To formalize the notion that M is a partial summary of D, we assume that (X, Y) and M are conditionally independent, given D:

$$P_{XY|D,M} = P_{XY|D} \tag{1}$$

This implies that, given D, the joint distribution of X and Y does not change with knowledge of M. An important special case where this holds in when $M = g(D)$ for some deterministic $g : \mathcal{D} \to \mathcal{M}$. We illustrate this probabilistic framework with motivating examples in Table 1, whose implications will be discussed throughout the paper.

Table 1: Examples of domains and metadata in different tasks.

Task	**Input** X	**Label** Y	**Domain** D	**Metadata** M
Sentiment annotation (Multiple Annotators)	Sentence to be annotated	Sentiment label (e.g., positive, negative)	Annotator identity (e.g., "Annotator 1")	Annotator's demographic profile (e.g., age)
Review rating prediction (Multiple Reviewers)	Product review text	Numerical rating (e.g., 1–5 stars)	Reviewer identity (e.g., "Reviewer 2")	Unlabeled texts written by the reviewer $\{X_i\}_{i=1}^n \overset{iid}{\sim} P_{X\|D=d}$
Image classification across styles	Image	Object category label (e.g., dog, car)	Image style (e.g., photograph, sketch, painting)	Textual description of style

The training data available to the learner is generated as follows: First, N domains $d_1, \ldots, d_N$ are sampled iid from P_D, but not observed. Then, conditioned on these d_i, corresponding values m_i are observed. In addition, for each i, $1 \leq i \leq N$, data (x_{ij}, y_{ij}) are sampled iid from $P_{XY|D=d_i}$, $1 \leq j \leq n_i$. In summary, the overall training data is

$$\left(m_i, (x_{ij}, y_{ij})_{j=1}^{n_i}\right)_{i=1}^{N}.$$

The goal of the learner is to produce a function f that accurately predicts labels on a new, random domain. In particular, f should minimize the risk

$$R(f) := \mathbb{E}_{X,Y,M,D}\left[\mathbb{1}_{f(\cdot)\neq Y}\right].$$

[2]Blanchard et al. [2011, 2021] use P_{XY} to denote a random domain, whereas in our notation, a random domain is either $P_{XY|D}$ or just D. Our introduction of D for a random domain allows us to use P_{XY} for the "average" domain, which will be a critical concept in what follows.

In practice, this risk is estimated by holding out several of the domains, and averaging the test errors on them. This probabilistic framing of DG generalizes that of Blanchard et al. [2011, 2021]. They focus on the special case where M is the marginal distribution of X for the given domain, and focus on the challenges associated to learning from empirical samples of the training and test X-marginals.

The training setup described above naturally gives rise to two different ways of using the available data. On one hand, the learner may choose to ignore the domain information and simply pool together all training samples, treating them as if they were drawn iid from a single distribution. On the other hand, the learner may choose to leverage the observed metadata m_i, which serves as side information about the underlying domain. These two strategies lead to two corresponding empirical risk minimization principles. Thus, let $\mathcal{F} \subset \{\mathcal{X} \times \mathcal{M} \to \mathcal{Y}\}$ denote a class of functions that take both input feature x and auxiliary metadata m as input, and $\mathcal{G} \subset \{\mathcal{X} \to \mathcal{Y}\}$ a class of functions that take only x as input. Consider two empirical risk minimizers:

$$\textbf{Pooling ERM: } \widehat{f}_{\text{pool}} = \arg\min_{f \in \mathcal{G}} \frac{1}{N} \sum_{i=1}^{N} \frac{1}{n_i} \sum_{j=1}^{n_i} \ell(y_{ij}, f(x_{ij})). \tag{2}$$

$$\textbf{Domain-informed (DI) ERM: } \widehat{f}_{\text{DG}} = \arg\min_{f \in \mathcal{F}} \frac{1}{N} \sum_{i=1}^{N} \frac{1}{n_i} \sum_{j=1}^{n_i} \ell(y_{ij}, f(x_{ij}, m_i)). \tag{3}$$

We are interested in when DI-ERM outperforms pooling ERM. From a theoretical perspective, we work in the large-sample and "large-model" limit (where $\mathcal{F}$ and $\mathcal{G}$ can approximate the Bayes-optimal predictor arbitrarily well). In this regime, standard learning-theoretic arguments imply that the performance of the two approaches is characterized by their corresponding Bayes risks, defined below.

4 Risk and Bayes risk in domain generalization

To aid in understanding domain generalization, it is helpful to consider DG in relation to two other problem settings. These settings differ only in what information the classifier f has access to. In all cases, the performance measure is the risk

$$R(f) := \mathbb{E}_{X,Y,M,D}\left[\mathbb{1}_{f(\cdot) \neq Y}\right],$$

where the argument of $f(\cdot)$ depends on settings.

No Domain Information: In this setting, the classifier only has access to the feature vector x at test time, and is thus $f(x)$. As noted earlier, most empirical DG methods, especially in computer vision, have this form. The risk in this case is

$$R(f) = \mathbb{E}_{X,Y,M,D}\left[\mathbb{1}_{f(X) \neq Y}\right] = \mathbb{E}_{X,Y}\left[\mathbb{1}_{f(X) \neq Y}\right] = \mathbb{E}_{X}\left[\mathbb{E}_{Y|X}\left[\mathbb{1}_{f(X) \neq Y}\right]\right],$$

where, because f does not depend on D or M, these variables marginalize out. Therefore, the problem reduces to learning with respect to the marginal distribution of (X, Y), which corresponds to pooling data across domains. The optimal classifier f^*_{pool} is thus the Bayes classifier for the marginal distribution of (X, Y):

$$f^*_{\text{pool}}(x) = \arg\max_{k} \mathbb{P}\left(Y = k | X = x\right).$$

The corresponding Bayes risk, R^*_{pool}, is the Bayes risk for the marginal distribution of (X, Y)

$$R^*_{\text{pool}} := \mathbb{E}_{X,Y}\left[\mathbb{1}_{f^*_{\text{pool}}(X) \neq Y}\right] = \mathbb{E}_{X}\left[1 - \max_{k} \mathbb{P}\left(Y = k | X\right)\right].$$

Full Domain Information: In this setting, the classifier has full knowledge of the domain D at test time, and is thus denoted $f(x, d)$. In practice, full knowledge of D is not available, and this setting therefore serves as a bound on the best possible performance of DG. The risk in this setting is

$$R(f) = \mathbb{E}_{X,Y,M,D}\left[\mathbb{1}_{f(X,D) \neq Y}\right] = \mathbb{E}_{X,Y,D}\left[\mathbb{1}_{f(X,D) \neq Y}\right] = \mathbb{E}_{X,D}\left[\mathbb{E}_{Y|X,D}\left[\mathbb{1}_{f(X,D) \neq Y}\right]\right].$$

The optimal classifier f^*_{full} is now the Bayes classifier for the distribution of $X, Y|D$,

$$f^*_{\text{full}}(x, d) = \arg\max_k \mathbb{P}\left(Y = k|X = x, D = d\right),$$

and the corresponding Bayes risk is:

$$R^*_{\text{full}} := \mathbb{E}_{X,Y,D}\left[\mathbb{1}_{f^*_{\text{full}}(X,D)\neq Y}\right] = \mathbb{E}_{X,D}\left[1 - \max_k \mathbb{P}\left(Y = k|X, D\right)\right].$$

In this setting, the classifier has full knowledge of the test domain, in other words, the joint distribution of (X, Y) for the given test domain. Therefore, $R^*(X, D)$ is the Bayes risk for the test domain, which serves as a lower bound for the risk in domain generalization.

Remark 1 *Achieving $R^*(X, D)$ is the goal of* domain adaptation.

Partial Domain Information: This is the setting of domain generalization. The classifier has access to not only x, but also a variable m that conveys partial information about the true domain d. A classifier in this setting is denoted $f(x, m)$. The risk is

$$R(f) = \mathbb{E}_{X,Y,M,D}\left[\mathbb{1}_{f(X,M)\neq Y}\right] = \mathbb{E}_{X,Y,M}\left[\mathbb{1}_{f(X,M)\neq Y}\right] = \mathbb{E}_{X,M}\left[\mathbb{E}_{Y|X,M}\left[\mathbb{1}_{f(X,M)\neq Y}\right]\right].$$

The optimal classifier f^*_{DG} is now the Bayes classifier for the distribution of $X, Y|M$,

$$f^*_{\text{DG}}(x, m) = \arg\max_k \mathbb{P}\left(Y = k|X = x, M = m\right),$$

and the corresponding Bayes risk is

$$R^*_{\text{DG}} := \mathbb{E}_{X,Y,M}\left[\mathbb{1}_{f^*_{\text{DG}}(X,M)\neq Y}\right] = \mathbb{E}_{X,M}\left[1 - \max_k \mathbb{P}\left(Y = k|X, M\right)\right].$$

In this setting, the optimal classifier uses both X and domain-specific signal M to predict Y. This setting aligns with the original theoretical motivations of DG and highlights the value of leveraging test-time domain information. A key goal of our work is to reassert the importance of this formulation and demonstrate both its theoretical advantages and empirical benefits, particularly in contrast to the more commonly used $f(x)$ setting.

5 Comparison of Bayes risks

This section develops bounds that relate the three Bayes risks defined in the previous section. The bounds reveal settings where domain information is and is not beneficial. The following basic result provides a starting point.

Proposition 1 (Risk Hierarchy) $R^*_{\text{pool}} \geq R^*_{\text{DG}} \geq R^*_{\text{full}}$.

The proof is straightforward (see Appendix C.1). The first inequality is trivial, as extending a feature vector can never decrease the Bayes risk. The second inequality follows from (1).

Our focus in this section is to determine conditions under which these inequalities become strict, and with a quantifiable gap. Toward that end, consider the following definition.

Definition 1 (Point-wise Margin) *Consider any random triple (X, Y, M), where Y is discrete. Define the point-wise margin of $Y|M = m, X = x$ as,*

$$\gamma(x, m) := \max_k \mathbb{P}\left(Y = k|X = x, M = m\right) - \text{2nd}\max_k \mathbb{P}\left(Y = k|X = x, M = m\right).$$

The operator $\text{2nd}\max_k$ returns the second largest value of its argument. Thus, if the two largest values of $\mathbb{P}\left(Y = k|X = x, M = m\right)$ are the same, $\gamma(x, m) = 0$. Intuitively, $\gamma(x, m)$ reflects the degree of certainty that the Bayes classifier $f^*_{\text{DG}}(x, m)$ has about its prediction. The larger $\gamma(x, m)$, the more confident the prediction.

The next result gives upper and lower bounds on the gap between R^*_{DG} and R^*_{pool}. This gap is the additional reduction in risk that results from leveraging the partial domain information M.

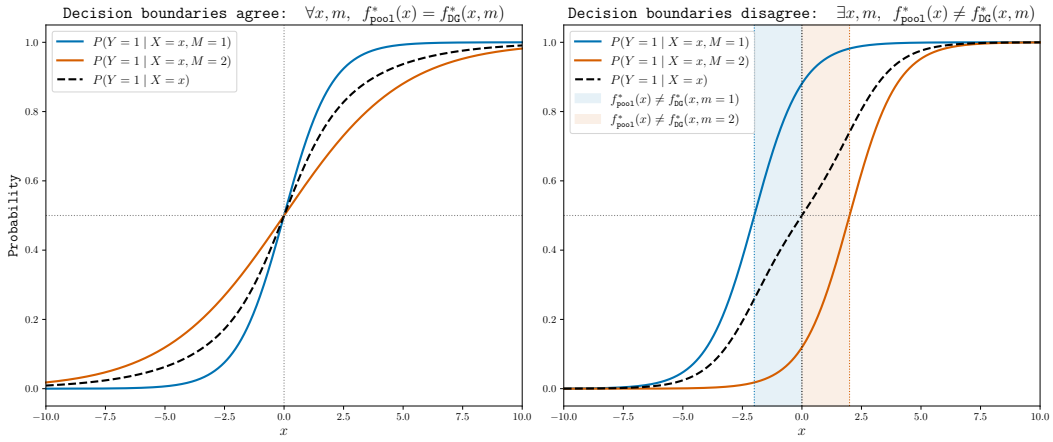

Figure 1: Illustration of Theorem 1. Consider binary classification with $X \in \mathbb{R}$, $Y \in \{1, 2\}$, and $M \in \{1, 2\}$. Then the Bayes classifiers $f^*_{\text{pool}}(x)$, $f^*_{\text{DG}}(x, m = 1)$ and $f^*_{\text{DG}}(x, m = 2)$ can be obtained by thresholding the corresponding posteriors at $1/2$. The left figure shows a scenario where the domain-informed classifier f^*_{DG} and the pooled classifier f^*_{pool} agree everywhere, and therefore both upper and lower bound are 0. In this case, domain information M is not beneficial. The right figure shows a scenario where f^*_{DG} disagrees with f^*_{pool} in certain regions, and domain information does lead to lower Bayes risk.

Theorem 1 (Risk Reduction from Domain Information) *Consider any random triple (X, Y, M), where Y is discrete. Then*

$$\mathbb{E}_{X,M}\left[\gamma(X, M)\mathbb{1}_{f^*_{\text{pool}}(X) \neq f^*_{\text{DG}}(X,M)}\right] \leq R^*_{\text{pool}} - R^*_{\text{DG}} \leq \mathbb{E}_{X,M}\left[\mathbb{1}_{f^*_{\text{pool}}(X) \neq f^*_{\text{DG}}(X,M)}\right].$$

The proof of Theorem 1 is in Appendix C.2. The upper bound represents the probability of disagreement between the domain-informed classifier f^*_{DG} and the pooled classifier f^*_{pool}. The lower bound can be interpreted as the expected cost of disagreement, where the cost is zero when the predictions agree, and equals the margin $\gamma(X, M)$ when they differ. Hence, domain information is particularly beneficial when f^*_{DG} frequently disagrees with f^*_{pool} in regions of high confidence. Figure 1 gives more intuition.

Remark 2 *Although not the focus of this paper, a version of Theorem 1 also holds for the gap $R^*_{\text{DG}} - R^*_{\text{full}}$, where R^*_{full} is the risk of a classifier that has full knowledge of the test domain. This bound quantifies the difference between DG and domain adaptation, and addresses a question left open by Blanchard et al. [2021, Lemma 9], see Appendix B for more detailed discussion.*

5.1 No information-theoretic gain under covariate shift

Theorem 1 holds regardless of the distribution P_{XYMD}. By considering assumptions on this distribution, stronger conclusions may be drawn. Covariate shift refers to the setting where, as the domain D varies, $P_{X|D}$ changes, but $P_{Y|X,D}$ does not. More generally, we can extend the meaning of covariate shift to be any DG problem where the Bayes classifier f^*_{DG} does not depend on M. In such a scenario, domain-specific information is of no benefit.

Corollary 1 *Under covariate shift, $R^*_{\text{pool}} = R^*_{\text{DG}}$.*

5.2 Information-theoretic gain under posterior drift

We now examine a class of distributions where the gap $R^*_{\text{pool}} - R^*_{\text{DG}}$ has a more concrete lower bound. This class of distributions is motivated by applications—particularly in natural language processing—where the posterior $P_{Y|X,D}$ differs across domains due to inherent ambiguity or subjectivity. A canonical example is the sentiment or toxicity annotation task, where annotators often disagree

on the same text. For instance, in the age-related sentiment analysis dataset of Díaz et al. [2018], the sentence *"Old people's appearance contains so much lived life."* received conflicting labels: 2/5 annotators seeing it as 'very positive', 2/5 as 'somewhat positive', and 1/5 as 'very negative'. This reflects how labeling tendency varies with annotator identity. We capture this phenomenon by introducing a formal posterior drift class for domain generalization.

Definition 2 (Posterior Drift Class for Domain Generalization)

$$\Pi(\gamma, \epsilon) := \Big\{ (X, Y, M, D) : \forall x, m,\ \gamma(x, m) \geq \gamma,\ \textit{and} \\ P_{X,M,M'}\Big(f^*_{\mathrm{DG}}\big(X, M\big) \neq f^*_{\mathrm{DG}}\big(X, M'\big)\Big) \geq \epsilon \Big\},$$

where $(M, M') \mid X \sim P_{M|X} \otimes P_{M|X}$ *are two independent draws.*

This class of DG problems captures settings where optimal classifiers with different M make conflicting predictions on a non-negligible region of the input space. The parameter γ quantifies the point-wise confidence of the optimal predictor, the parameter ϵ quantifies the average amount of variation in $P_{Y|X,M}$ for different M. With this, we have an explicit lower bound:

Proposition 2

$$\inf_{(X,Y,M)\in\Pi(\gamma,\epsilon)} \Big[R^*_{\mathrm{pool}} - R^*_{\mathrm{DG}}\Big] \geq \frac{\gamma \cdot \epsilon}{2}$$

The proof is in Appendix C.3. This lower bound shows that leveraging domain-specific information yields a provable benefit of at least $\gamma\epsilon/2$ for this particular formulation of posterior drift.

In contrast to pessimistic results in *domain adaptation*—where no method consistently outperforms vanilla ERM under posterior drift [Zhu et al., 2024, Liu et al., 2024]—our work presents an optimistic view in *domain generalization*: by conditioning on domain metadata M, we can provably do better than pooling-based prediction.

5.3 Advantage of DI-ERM beyond posterior drift

The previous subsection demonstrates the information-theoretic gain from incorporating domain information M. We now consider the practical setting of learning under restricted function classes.

Let $\mathcal{F} \subset \{\mathcal{X} \times \mathcal{M} \to \mathcal{Y}\}$ denote a class of predictors that take both input features x and auxiliary metadata m as input. From this class, we define a corresponding class $\mathcal{G} \subset \{\mathcal{X} \to \mathcal{Y}\}$ as:

$$\mathcal{G} := \{x \mapsto f(x, m_0) : f \in \mathcal{F},\ m_0 \in \mathcal{M}\},$$

i.e., $\mathcal{G}$ consists of classifiers in $\mathcal{F}$ where the metadata variable m is held fixed.

Clearly, any function in $\mathcal{G}$ is realizable within $\mathcal{F}$. Therefore,

$$R^*_{\mathrm{pool},\,\mathcal{G}} := \inf_{f\in\mathcal{G}} R(f) \quad \geq \quad R^*_{\mathrm{DG},\,\mathcal{F}} := \inf_{f\in\mathcal{F}} R(f). \tag{4}$$

We are interested in understanding when strict inequality holds. The example below show that even if there is no information-theoretic gain of DI-ERM under covariate shift (Corollary 1), it may still have practical advantage when considering a restricted function class $\mathcal{F}$.

Example 1 (Covariate shift without posterior drift) *Let* P_{XYM} *be*

$$M \sim \mathrm{Bernoulli}(p),\ \textit{where } p > 1/2, \quad \begin{cases} M = 1: & X \sim \mathrm{Unif}[0, 2],\ Y = \mathrm{sign}(X - 1) \\ M = 2: & X \sim \mathrm{Unif}[4, 6],\ Y = \mathrm{sign}(X - 5). \end{cases}$$

Because the supports are disjoint, the pooling and DG Bayes classifier are the same, to be specific

$$f^*_{\mathrm{pool}}(x) = \begin{cases} \mathrm{sign}(x - 1), & x \in [0, 2] \\ \mathrm{sign}(x - 5), & x \in [4, 6] \end{cases}, \quad f^*_{\mathrm{DG}}(x, m) = \mathrm{sign}(x - 4m + 3) \quad \Longrightarrow f^*_{\mathrm{pool}} = f^*_{\mathrm{DG}}$$

therefore $R^*_{\mathrm{pool}} = R^*_{\mathrm{DG}} = 0$. The model classes are linear classifiers

$$\mathcal{F} = \big\{f(x,m) = \mathrm{sign}(w^\top x + v^\top m + b)\big\}, \qquad \mathcal{G} = \big\{f(x) = \mathrm{sign}(w^\top x + b)\big\}.$$

$\mathcal{F}$ can realize f^*_{DG} with a bias term that depends on m, giving $R^*_{\mathrm{DG},\,\mathcal{F}} = 0$. However, a predictor in $\mathcal{G}$ can only choose a single threshold, and the optimal one is

$$f^*_{\mathrm{pool},\,\mathcal{G}}(x) = \mathrm{sign}(x-1), \quad R^*_{\mathrm{pool},\,\mathcal{G}} = \frac{\min\{p, 1-p\}}{2}.$$

Therefore, $R^*_{\mathrm{pool},\,\mathcal{G}} > R^*_{\mathrm{DG},\,\mathcal{F}}$, even though $R^*_{\mathrm{pool}} = R^*_{\mathrm{DG}}$.

This toy construction mirrors image classification task across different styles: each style (domain) has a separate support, so the Bayes classifier is the same with or without m, yet m still helps within a restricted model class. This is experimentally verified in Section 6 .

6 Experiments

We evaluate the effectiveness of domain-informed ERM (DI-ERM) in three experimental settings. Our primary focus is on the comparison between DI-ERM and pooling ERM, which highlights the benefit of incorporating domain metadata. Additional results—including linear probing, benchmarks against alternative methods, and complete experimental details—are provided in Appendix D.

Sentiment disagreement among annotators In many NLP tasks, annotators exhibit subjective preferences, leading to disagreement of the label y on the same input x—a form of posterior drift discussed in Section 5.2. To study this phenomenon, we use the dataset of Díaz et al. [2018], which re-annotates a subset of Sentiment140 for training and provides a test set drawn from age-related blog posts. The training set comprises 59,235 sentences labeled by 1,481 annotators; the test set includes 1,419 sentences labeled by 878 annotators. Each sentence is annotated by 4–5 individuals, and the labels exhibit high disagreement (about 40 $\%$). In this setting, the input x is a sentence, the label $y \in \{1,2,3,4,5\}$ denotes sentiment on a five-point scale, the domain d corresponds to the annotator, and the domain information m consists of demographic metadata (e.g., age, upbringing region).

To encode domain information M, we concatenate it with the sentence x in a text-prompt format, as illustrated in Figure 2. Table 3 reports the results. DI-ERM substantially outperforms pooling ERM, demonstrating that leveraging annotator metadata can dramatically improve predictive accuracy. Notably, DI-ERM also surpasses the previous state-of-the-art reported by Deng et al. [2023].

Table 3: Test accuracy on the sentiment disagreement dataset. Incorporating annotator profiles (M) through DI-ERM yields a dramatic improvement over pooling ERM, reflecting the importance of modeling annotator-specific posterior drift. In particular, DI-ERM nearly doubles accuracy compared to pooling ERM and surpasses the previous state-of-the-art (69.8% by [Deng et al., 2023]).

Algorithm	Model	Test Avg Acc
Pooling ERM (finetune)	BERT	49.1 ± 0.4
DI-ERM (finetune)	BERT	90.5 ± 0.2

Reviewer-specific sentiment analysis We next examine the WILDS-Amazon Reviews dataset [Koh et al., 2021], which captures distributional shifts across reviewers. Here, the input x is a product review, $y \in \{1,\dots,5\}$ is the star rating, d denotes the reviewer identity, and m consists of all (unlabeled) reviews written by that reviewer.

The central hypothesis is that a reviewer's writing style $M = P_{X|D}$ provides a useful signal for predicting their rating behavior $P_{Y|X,D}$. The training set contains 245,502 reviews from 1,252 reviewers, while the test set consists of 100,050 reviews from 1,334 unseen reviewers.

To incorporate reviewer context M, we randomly sample 20 additional reviews written by the same reviewer and concatenate them with the current review in a prompt format, shown in Figure 3. As summarized in Table 4, DI-ERM outperforms pooling ERM. Beyond higher average accuracy, DI-ERM also boosts the 10th-percentile accuracy across reviewers—a key robustness metric used on the official leaderboard. With end-to-end fine-tuning, DI-ERM surpassing the best leaderboard result on both metrics (see Appendix D for additional discussion).

Table 4: Sentiment classification performance on Amazon-WILDS with reviewer-specific context. DI-ERM consistently improves over pooling ERM, both in average accuracy and in 10th-percentile reviewer accuracy—the official leaderboard metric. It also exceeds the best result reported on the WILDS leaderboard (`https://wilds.stanford.edu/`).

Algorithm	Model	Test Avg Acc	Test 10% Acc
Pooling ERM (finetune)	nomic-embed-text-v1.5	71.8 ± 0.9	54.7 ± 0.0
DI-ERM (finetune)	nomic-embed-text-v1.5	73.1 ± 0.3	56.4 ± 0.8

Image classification across styles We next evaluate our method on the PACS dataset [Li et al., 2017], which contains images drawn from four distinct visual styles: $d \in$ Photo (P), Art Painting (A), Cartoon (C), Sketch (S). Each image x belongs to one of seven categories, $y \in$ {Dog, Elephant, Giraffe, Guitar, Horse, House, Person} . Domain information is represented by a short text description m, such as "a photo" or "a pencil sketch" (see Figure 4).

This vision task satisfies covariate shift, since a single classifier should accurately classify all images across domains. Thus, in line with Section 5.3, we expect any gains to be due to using a restricted function class. To implement DI-ERM, we use pretrained image foundation models (e.g., CLIP [Radford et al., 2021], DINOv2 [Oquab et al., 2023]) to extract visual features from x, and encode the domain description m using a pretrained language model (DistilBERT) following the prompt in Figure 4. The resulting image and text embeddings are concatenated into a joint representation for classification.

We follow the standard PACS evaluation protocol: training on three domains and testing on the held-out fourth domain, repeated across all domain splits. All encoders are frozen, and linear classifiers are trained on top of the fixed representations.

As shown in Table 10, DI-ERM improves over pooling ERM in most settings. The gains are most pronounced for mid-sized models, while the benefit diminishes for larger foundation models. This pattern aligns with the discussion in Section 5.3: under covariate shift, the benefit of DI-ERM decreases as model mismatch becomes smaller.

Table 5: Domain generalization results on PACS using models from the CLIP and DINOv2 families. DI-ERM achieves improved accuracy over pooling ERM in most configurations, particularly for mid-sized models, less noticeable for large-sized models. When using large-sized models, both ERM and DI-ERM approaches SOTA performance.

Algorithm	Model	PAC $\to$ S	ACS $\to$ P	CSP $\to$ A	SPA $\to$ C	Test Avg Acc
Pooling ERM (linear)	CLIP: vitb32	86.97	99.58	95.90	97.48	94.98
DI-ERM (linear)		**88.06**	**99.64**	**96.29**	**97.48**	**95.37**
Pooling ERM (linear)	CLIP: vitl14	95.42	99.94	99.22	99.79	98.59
DI-ERM (linear)		95.32	99.94	**99.32**	99.79	98.59
Pooling ERM (linear)	DINOv2: vits14	79.82	85.81	93.55	91.34	87.63
DI-ERM (linear)		**80.45**	**90.00**	**94.09**	**91.60**	**89.04**
Pooling ERM (linear)	DINOv2: vitl14	92.29	96.41	98.14	97.48	96.08
DI-ERM (linear)		**92.42**	**97.37**	98.10	97.48	**96.34**

7 Conclusions

This work presents a rigorous theory of domain generalization, precisely characterizing when and why leveraging domain information at test time is beneficial. Empirically, we demonstrate that domain-informed ERM (DI-ERM) outperforms pooled ERM across three representative scenarios in language and vision tasks. Future work can be done to explore alternative ways of encoding domain information, and a broader range of DG benchmarks.

References

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

Ruqi Bai, Saurabh Bagchi, and David I. Inouye. Benchmarking algorithms for federated domain generalization. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=wprSv7ichW.

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

Gilles Blanchard, Gyemin Lee, and Clayton Scott. Generalizing from several related classification tasks to a new unlabeled sample. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc., 2011. URL https://proceedings.neurips.cc/paper_files/paper/2011/file/b571ecea16a9824023ee1af16897a582-Paper.pdf.

Gilles Blanchard, Aniket Anand Deshmukh, Urun Dogan, Gyemin Lee, and Clayton Scott. Domain generalization by marginal transfer learning. *Journal of machine learning research*, 22(2):1–55, 2021.

T Tony Cai and Hongji Wei. Transfer learning for nonparametric classification: Minimax rate and adaptive classifier. *The Annals of Statistics*, 49(1):100–128, 2021.

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

Marie-Catherine De Marneffe, Mandy Simons, and Judith Tonhauser. The commitmentbank: Investigating projection in naturally occurring discourse. In *proceedings of Sinn und Bedeutung*, volume 23, pages 107–124, 2019.

Naihao Deng, Xinliang Zhang, Siyang Liu, Winston Wu, Lu Wang, and Rada Mihalcea. You are what you annotate: Towards better models through annotator representations. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12475–12498, 2023.

Aniket Anand Deshmukh, Ankit Bansal, and Akash Rastogi. Domain2vec: Deep domain generalization. *arXiv preprint arXiv:1807.02919*, 2018.

Mark Díaz, Isaac Johnson, Amanda Lazar, Anne Marie Piper, and Darren Gergle. Addressing age-related bias in sentiment analysis. In *Proceedings of the 2018 chi conference on human factors in computing systems*, pages 1–14, 2018.

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*, pages 1657–1664, 2013.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario March, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of machine learning research*, 17(59):1–35, 2016.

Henry Gouk, Ondrej Bohdal, Da Li, and Timothy Hospedales. On the limitations of general purpose domain generalisation methods, 2024. URL https://arxiv.org/abs/2202.00563.

Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. In *International Conference on Learning Representations*, 2021.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International conference on machine learning*, pages 5637–5664. PMLR, 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*, pages 5542–5550, 2017.

Jiashuo Liu, Tianyu Wang, Peng Cui, and Hongseok Namkoong. Rethinking distribution shifts: Empirical analysis and inductive modeling for tabular data, 2024. URL https://arxiv.org/abs/2307.05284.

Subha Maity, Diptavo Dutta, Jonathan Terhorst, Yuekai Sun, and Moulinath Banerjee. A linear adjustment based approach to posterior drift in transfer learning. *Biometrika*, 2024. URL https://doi.org/10.1093/biomet/asad029.

Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh. Domain adaptation: Learning bounds and algorithms. In *Proceedings of The 22nd Annual Conference on Learning Theory (COLT 2009)*, Montréal, Canada, 2009. URL http://www.cs.nyu.edu/~mohri/postscript/nadap.pdf.

Krikamol Muandet, David Balduzzi, and Bernhard Schölkopf. Domain generalization via invariant feature representation. In *International conference on machine learning*, pages 10–18. PMLR, 2013.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2023.

Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1406–1415, 2019.

Barbara Plank. The "problem" of human label variation: On ground truth in data, modeling and evaluation. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10671–10682, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.731. URL https://aclanthology.org/2022.emnlp-main.731/.

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*, pages 8748–8763. PmLR, 2021.

Elan Rosenfeld, Pradeep Ravikumar, and Andrej Risteski. The risks of invariant risk minimization. In *International Conference on Learning Representations*, volume 9, 2021.

Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=ryxGuJrFvS.

Shiori Sagawa, Pang Wei Koh, Tony Lee, Irena Gao, Sang Michael Xie, Kendrick Shen, Ananya Kumar, Weihua Hu, Michihiro Yasunaga, Henrik Marklund, Sara Beery, Etienne David, Ian Stavness, Wei Guo, Jure Leskovec, Kate Saenko, Tatsunori Hashimoto, Sergey Levine, Chelsea Finn, and Percy Liang. Extending the WILDS benchmark for unsupervised adaptation. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=z7p2V6KROOV.

Clayton Scott. A generalized Neyman-Pearson criterion for optimal domain adaptation. In *Algorithmic Learning Theory*, pages 738–761. PMLR, 2019.

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

Piotr Teterwak, Kuniaki Saito, Theodoros Tsiligkaridis, Kate Saenko, and Bryan Plummer. Erm++: An improved baseline for domain generalization. In *Proceedings of the Winter Conference on Applications of Computer Vision (WACV)*, pages 8514–8524, February 2025.

Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5018–5027, 2017.

Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. In *Advances in Neural Information Processing Systems*, pages 10506–10518, 2019.

Jindong Wang, Cuiling Lan, Chang Liu, Yidong Ouyang, Tao Qin, Wang Lu, Yiqiang Chen, Wenjun Zeng, and S Yu Philip. Generalizing to unseen domains: A survey on domain generalization. *IEEE transactions on knowledge and data engineering*, 35(8):8052–8072, 2022.

Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Russ R Salakhutdinov, and Alexander J Smola. Deep sets. *Advances in neural information processing systems*, 30, 2017.

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

Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. Domain generalization: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4396–4415, 2023.

Yilun Zhu, Jianxin Zhang, Aditya Gangrade, and Clayton Scott. Label noise: Ignorance is bliss. *Advances in Neural Information Processing Systems*, 38:116575–116616, 2024.

A Appendix: additional literature review

ERM is hard to beat. Empirically, Gulrajani and Lopez-Paz [2021] first emphasized that a well-tuned empirical risk minimization (ERM) baseline outperforms many domain generalization (DG) methods on vision benchmarks. Similar patterns were later observed on the WILDS benchmark [Koh et al., 2021], and again in the context of federated learning by Bai et al. [2024]. On the theoretical front, Rosenfeld et al. [2021] and Gouk et al. [2024] studied function classes of the form $f : \mathcal{X} \to \mathcal{Y}$ and concluded that, under common assumptions, ERM cannot be fundamentally outperformed (e.g., in terms of minimax risk).

The use of unlabeled data. While most DG methods restrict themselves to using only the input x at inference time, some methods explore the use of unlabeled test-domain data. Several DG methods attempt to exploit unlabeled test data to improve generalization [Blanchard et al., 2011, Muandet et al., 2013, Zhang et al., 2021]. A closely related setting is unsupervised domain adaptation (UDA), where unlabeled test data are used to adapt models to the test domain. Unlike DG, UDA assumes access to target-domain data at training time and typically requires learning a separate model per test domain [Sun and Saenko, 2016, Ganin et al., 2016].

Although promising in principle, the practical benefits of using unlabeled data remain mixed. A large-scale study by Sagawa et al. [2022] evaluating methods across ten diverse datasets found that incorporating unlabeled data frequently failed to improve upon strong ERM baselines. These findings reinforce the need for a more precise understanding of when and how unlabeled data can contribute to domain generalization.

Our framework addresses this gap by casting unlabeled data as a special case of auxiliary domain information, and by providing conditions under which such information is expected to improve generalization performance.

B Appendix: Partial versus full domain knowledge

Although not the focus of this paper, a version of Theorem 1 also holds for the gap $R^*_{\mathrm{DG}} - R^*_{\mathrm{full}}$, where R^*_{full} is the risk of a classifier that has full knowledge of the test domain. Such a bound addresses a question left open by Blanchard et al. [2021, Lemma 9], who established that this gap is lower bounded by zero, and provide a condition under which the gap equals zero. The following result bounds this gap in a more general setting.

Proposition 3 *Let*

$$\widetilde{\gamma}(x,d) := \max_k \mathbb{P}\left(Y=k|X=x,D=d\right) - \operatorname{2nd}\max_k \mathbb{P}\left(Y=k|X=x,D=d\right).$$

Then

$$\mathbb{E}_{X,D,M}\left[\widetilde{\gamma}(X,D)\mathbb{1}_{f^*_{\mathrm{full}}(X,D)\neq f^*_{\mathrm{DG}}(X,M)}\right] \leq R^*_{\mathrm{DG}} - R^*_{\mathrm{full}} \leq \mathbb{E}_{X,D,M}\left[\mathbb{1}_{f^*_{\mathrm{full}}(X,D)\neq f^*_{\mathrm{DG}}(X,M)}\right]$$

and in particular,

$$f^*_{\mathrm{full}}(x,d) = f^*_{\mathrm{DG}}(x,m) \quad \textit{almost surely w.r.t. } P_{XMD} \implies R^*_{\mathrm{full}} = R^*_{\mathrm{DG}}.$$

The result has an interpretation analogous to that of Theorem 1. In particular, if the observed domain information is of low quality, in the sense that f^*_{full} disagrees with f^*_{DG} often, and with high confidence, then R^*_{DG} can be substantially worse than R^*_{full}.

C Appendix: proofs

C.1 Proof of Proposition 1

Proposition (Risk Hierarchy) $R^*_{\mathrm{pool}} \geq R^*_{\mathrm{DG}} \geq R^*_{\mathrm{full}}$.

Proof.

$$
\begin{aligned}
R^*_{\text{pool}} &= \inf_{f:\mathcal{X}\to\mathcal{Y}} \mathbb{E}_{X,Y,M,D}\left[\mathbb{1}_{f(X)\neq Y}\right] \\
&\geq \inf_{f:\mathcal{X}\times\mathcal{M}\to\mathcal{Y}} \mathbb{E}_{X,Y,M,D}\left[\mathbb{1}_{f(X,M)\neq Y}\right] = R^*_{\text{DG}} \\
&\geq \inf_{f:\mathcal{X}\times\mathcal{M}\times\mathcal{D}\to\mathcal{Y}} \mathbb{E}_{X,Y,M,D}\left[\mathbb{1}_{f(X,M,D)\neq Y}\right] \\
&= \inf_{f:\mathcal{X}\times\mathcal{D}\to\mathcal{Y}} \mathbb{E}_{X,Y,M,D}\left[\mathbb{1}_{f(X,D)\neq Y}\right] = R^*_{\text{full}} \qquad \because (X,Y)|M,D = (X,Y)|D
\end{aligned}
$$

■

C.2 Proof of Theorem 1

Theorem (Risk Reduction from Domain Information) *Consider any random triple (X, Y, M), where Y is discrete. Then*

$$
\mathbb{E}_{X,M}\left[\gamma(X,M)\mathbb{1}_{f^*_{\text{pool}}(X)\neq f^*_{\text{DG}}(X,M)}\right] \leq R^*_{\text{pool}} - R^*_{\text{DG}} \leq \mathbb{E}_{X,M}\left[\mathbb{1}_{f^*_{\text{pool}}(X)\neq f^*_{\text{DG}}(X,M)}\right].
$$

Proof. The gap in the two risks can be expressed as

$$
\begin{aligned}
&R^*_{\text{pool}} - R^*_{\text{DG}} \\
&\quad= \mathbb{E}_X\left[\mathbb{E}_{M|X}\left[\mathbb{P}\left(Y = f^*_{\text{DG}}(X,M)|X,M\right)\right]\right] - \mathbb{E}_X\left[\mathbb{P}\left(Y = f^*_{\text{pool}}(X)|X\right)\right] \\
&\quad= \mathbb{E}_X\left[\mathbb{E}_{M|X}\left[\mathbb{P}\left(Y = f^*_{\text{DG}}(X,M)|X,M\right)\right]\right] - \mathbb{E}_X\left[\mathbb{E}_{M|X}\left[\mathbb{P}\left(Y = f^*_{\text{pool}}(X)|X,M\right)\right]\right] \\
&\quad= \mathbb{E}_X\left[\mathbb{E}_{M|X}\left[\mathbb{P}\left(Y = f^*_{\text{DG}}(X,M)|X,M\right) - \mathbb{P}\left(Y = f^*_{\text{pool}}(X)|X,M\right)\right]\right]
\end{aligned}
$$

Notice that for any x, m, if $f^*_{\text{DG}}(x,m) = f^*_{\text{pool}}(x)$, then the pointwise difference of the conditional probabilities inside the expectation above must be zero.

Whereas if they disagree, then it must hold that

$$
P(Y = f^*_{\text{pool}}(X)|X,M) \leq 2\text{nd}\max_k P(Y = k|X,M)
$$

due to the definition of f^*_{DG}.

It thus follows that.

$$
\begin{aligned}
\gamma(x,m)\mathbb{1}_{f^*_{\text{pool}}(x)\neq f^*_{\text{DG}}(x,m)} &\leq \mathbb{P}\left(Y = f^*_{\text{DG}}(x,m)|X=x,M=m\right) - \mathbb{P}\left(Y = f^*_{\text{pool}}(x)|X=x,M=m\right) \\
&\leq \mathbb{1}_{f^*_{\text{pool}}(x)\neq f^*_{\text{DG}}(x,m)}.
\end{aligned}
$$

The inequalities in the theorem statement now follow.

■

C.3 Proof of Proposition 2

Proposition

$$
\inf_{(X,Y,M)\in\Pi(\gamma,\epsilon)}\left[R^*_{\text{pool}} - R^*_{\text{DG}}\right] \geq \frac{\gamma\cdot\epsilon}{2}
$$

Proof. From the lower bound in Theorem 1, we have

$$
\begin{aligned}
R^*_{\text{pool}} - R^*_{\text{DG}} &\geq \mathbb{E}_{X,M}\left[\gamma(X,M)\mathbb{1}_{f^*_{\text{pool}}(X)\neq f^*_{\text{DG}}(X,M)}\right] && \because \text{ Theorem 1} \\
&\geq \gamma\,\mathbb{E}_{X,M}\left[\mathbb{1}_{f^*_{\text{pool}}(X)\neq f^*_{\text{DG}}(X,M)}\right] && \because \text{ margin assumption in } \Pi(\gamma,\epsilon) \\
&= \gamma\,\mathbb{E}_X\left[\mathbb{E}_{M|X}\left[\mathbb{1}_{f^*_{\text{pool}}(X)\neq f^*_{\text{DG}}(X,M)}\right]\right]
\end{aligned}
$$

Now we will show that

$$\forall x, \quad \mathbb{E}_{M|X=x}\left[\mathbb{1}_{f^*_{\text{pool}}(x)\neq f^*_{\text{DG}}(x,M)}\right] \geq \frac{1}{2}\,\mathbb{E}_{M,M'|X=x}\left[\mathbb{1}_{f^*_{\text{DG}}(x,M)\neq f^*_{\text{DG}}(x,M')}\right],$$

where

$$M,\ M' \overset{\text{i.i.d.}}{\sim} P_{M|X=x}.$$

Let's examine the two terms. Fix x, denote

$$\pi_k(x) = \mathbb{P}\left(f^*_{\text{DG}}(x,M) = k | X = x\right),$$

note that the randomness comes from M.

Then for any x,

$$\begin{aligned}
\mathbb{E}_{M,M'|X=x}\left[\mathbb{1}_{f^*_{\text{DG}}(x,M)\neq f^*_{\text{DG}}(x,M')}\right] &= \mathbb{P}\left(f^*_{\text{DG}}(x,M) \neq f^*_{\text{DG}}(x,M') | X = x\right) \\
&= \sum_k \mathbb{P}\left(f^*_{\text{DG}}(x,M) = k, f^*_{\text{DG}}(x,M') \neq k | X = x\right) \\
&= \sum_k \pi_k(x)\left(1 - \pi_k(x)\right) \\
&= 1 - \sum_k \pi_k(x)^2.
\end{aligned}$$

Now assume $f^*_{\text{pool}}(x) = k_0$, then

$$\begin{aligned}
\mathbb{E}_{M|X=x}\left[\mathbb{1}_{f^*_{\text{pool}}(x)\neq f^*_{\text{DG}}(x,M)}\right] &= \mathbb{E}_{M|X=x}\left[\mathbb{1}_{f^*_{\text{DG}}(x,M)\neq k_0}\right] \\
&= 1 - \pi_{k_0}(x)
\end{aligned}$$

Notice that

$$\begin{aligned}
1 - \sum_k \pi_k(x)^2 &\leq 1 - \pi_{k_0}^2 && \text{“=” when } \pi_{k_0} = 1 \\
&= (1 + \pi_{k_0})(1 - \pi_{k_0}) \\
&\leq 2(1 - \pi_{k_0}) && \text{“=” when } \pi_{k_0} = 1.
\end{aligned}$$

Then

$$\mathbb{E}_{M|X=x}\left[\mathbb{1}_{f^*_{\text{pool}}(x)\neq f^*_{\text{DG}}(x,M)}\right] \geq \frac{1}{2}\,\mathbb{E}_{M,M'|X=x}\left[\mathbb{1}_{f^*_{\text{DG}}(x,M)\neq f^*_{\text{DG}}(x,M')}\right].$$

Integrate over x, we have

$$\begin{aligned}
\mathbb{E}_X\left[\mathbb{E}_{M|X}\left[\mathbb{1}_{f^*_{\text{pool}}(x)\neq f^*_{\text{DG}}(X,M)}\right]\right] &\geq \frac{1}{2}\mathbb{E}_X\left[\mathbb{E}_{M,M'|X}\left[\mathbb{1}_{f^*_{\text{DG}}(X,M)\neq f^*_{\text{DG}}(X,M')}\right]\right] \\
&= \frac{1}{2}P_{X,M,M'}\Big(f^*_{\text{DG}}\big(X,M\big) \neq f^*_{\text{DG}}\big(X,M'\big)\Big) \\
&\geq \frac{1}{2}\epsilon \qquad \because \text{by definition of } \Pi(\gamma,\epsilon)
\end{aligned}$$

■

C.4 Proof of Proposition 3

Proposition *Let*

$$\widetilde{\gamma}(x,d) := \max_k \mathbb{P}\left(Y = k | X = x, D = d\right) - 2\text{nd}\max_k \mathbb{P}\left(Y = k | X = x, D = d\right)$$

Then

$$\mathbb{E}_{X,D,M}\left[\gamma(X,D)\mathbb{1}_{f^*_{\text{full}}(X,D)\neq f^*_{\text{DG}}(X,M)}\right] \leq R^*_{\text{DG}} - R^*_{\text{full}} \leq \mathbb{E}_{X,D,M}\left[\mathbb{1}_{f^*_{\text{full}}(X,D)\neq f^*_{\text{DG}}(X,M)}\right]$$

and in particular,

$$f^*_{\text{full}}(x,d) = f^*_{\text{DG}}(x,m) \quad \textit{almost surely w.r.t. } P_{XMD} \implies R^*_{\text{full}} = R^*_{\text{DG}}.$$

Proof.

$$\begin{aligned} R^*_{\text{DG}} - R^*_{\text{full}} &= \mathbb{E}_{X,Y,D,M}\left[\mathbb{1}_{Y \neq f^*_{\text{full}}(X,D)}\right] - \mathbb{E}_{X,Y,D,M}\left[\mathbb{1}_{Y \neq f^*_{\text{DG}}(X,M)}\right] \\ &= \mathbb{E}_{X,D,M}\left[\mathbb{P}\left(Y = f^*_{\text{DG}}(X,M)\right) - \mathbb{P}\left(Y = f^*_{\text{full}}(X,D)\right) | X, D, M\right] \end{aligned}$$

Recall the assumption on M:

$$Y|X,D,M = Y|X,D.$$

Then for every x, d and m,

$$\begin{aligned} &\mathbb{P}\left(Y = f^*_{\text{DG}}(x,m)|X=x,D=d,M=m\right) - \mathbb{P}\left(Y = f^*_{\text{full}}(x,d)|X=x,D=d,M=m\right) \\ &= \mathbb{P}\left(Y = f^*_{\text{DG}}(x,m)|X=x,D=d\right) - \mathbb{P}\left(Y = f^*_{\text{full}}(x,d)|X=x,D=d\right) \\ &\geq \gamma(x,d)\mathbb{1}_{f^*(x,m)\neq f^*(x,d)}. \end{aligned}$$

Similarily,

$$\begin{aligned} &\mathbb{P}\left(Y = f^*_{\text{DG}}(x,m)|X=x,D=d,M=m\right) - \mathbb{P}\left(Y = f^*_{\text{full}}(x,d)|X=x,D=d,M=m\right) \\ &\leq \mathbb{1}_{f^*(x,m)\neq f^*(x,d)}. \end{aligned}$$

Integrate over x, d, m, we get the lower and upper bound.

From the lower and upper bound, we can directly get the sufficient condition

$$f^*_{\text{full}}(x,d) = f^*_{\text{DG}}(x,m) \quad \text{almost surely w.r.t. } P_{XMD}. \implies R^*_{\text{full}} = R^*_{\text{DG}}$$

■

D Appendix: Experimental details

This section provides additional details on our experimental setup, models, and performance comparisons. Unless otherwise specified, all models used for fine-tuning are implemented using publicly available checkpoints (e.g., via Huggingface, Pytorch, or official Github repo). For linear probing experiments, we extract feature representations using pre-trained transformers and train downstream classifiers with `scikit-learn`, using either logistic regression or multilayer perceptrons (MLPs).

The following subsections follows the same structure as Section 6, while providing additional details and full tables.

D.1 Sentiment disagreement among annotators

Fine-Tuning. We fine-tune the `bert-base-uncased` model and benchmark DI-ERM against other domain generalization methods. For DI-ERM, we concatenate the sentence x with the annotator profile m using the text prompt shown in Figure 2.

Table 6 reports the results. Our models consistently outperform prior work, with the best configuration achieving over 90% test accuracy—substantially higher than the previous state-of-the-art reported by Deng et al. [2023].

Linear/MLP-probing. We also evaluate in a frozen-feature setting, where the language model is fixed and a lightweight classifier is trained on top. Here, x is encoded with a pretrained sentiment model (e.g., [CLS] embedding of DistilBERT fine-tuned on SST-2), while m is encoded with a general-purpose DistilBERT. The embeddings are concatenated and passed to either a linear or shallow MLP classifier. The classifiers are trained in `scikit-learn`.

Table 7 presents the results. DI-ERM consistently outperforms pooling ERM across different feature extractors.

```
Instruction: Read the following sentence and the annotator's
demographic profile and determine how positive or negative the
annotator judged the sentence on a 1-5 scale (1 = Very negative, 5
= Very positive).

Sentence: [sentence goes here]

Annotator profile: Age {age}, Race {race}, Hispanic/Latino {hisp},
grew up in {grew}, currently lives in {curr}, region {region}, income
{income}, education {education}, employment {employment}, living
situation {living}, politics {politics}, gender {gender}.

Answer:
```

Figure 2: Text prompt that encodes annotator profile.

Table 6: Test accuracy on the sentiment disagreement dataset (fine-tuning BERT). DI-ERM (ours) achieves the best performance.

Algorithm	**Model**	**Test Avg Acc**
ERM	BERT	49.1 ± 0.4
IRM	BERT	48.1 ± 0.7
GroupDRO	BERT	49.1 ± 0.1
CORAL	BERT	48.4 ± 0.2
AnnEmb (SOTA)	BERT	64.6 ± 0.8
DI-ERM (ours, fine-tune)	BERT	$\mathbf{90.5 \pm 0.2}$

Table 7: Test accuracy on the sentiment disagreement dataset (frozen feature extractor). DI-ERM consistently outperforms pooling ERM, and in some settings surpasses the prior state-of-the-art of Deng et al. [2023]. We highlight the best performance reported by Deng et al. [2023] (69.77) and our highest score (83.41). †: Checkpoints used in Deng et al. [2023] were not publicly specified.

Algorithm	**Model**	**Test Avg Acc**
Deng et al. [2023]	BERT†	64.61
	RoBERTa†	60.30
	DeBERTa†	69.77
Pooling ERM (linear)	distilbert-base-uncased-finetuned-sst-2-english	45.85
DI-ERM (linear)	distilbert-base-uncased-finetuned-sst-2-english	**46.42**
Pooling ERM (MLP)	distilbert-base-uncased-finetuned-sst-2-english	55.07
DI-ERM (MLP)	distilbert-base-uncased-finetuned-sst-2-english	**78.45**
Pooling ERM (linear)	bert-base-multilingual-uncased-sentiment	43.06
DI-ERM (linear)	bert-base-multilingual-uncased-sentiment	**43.94**
Pooling ERM (MLP)	bert-base-multilingual-uncased-sentiment	53.90
DI-ERM (MLP)	bert-base-multilingual-uncased-sentiment	**83.41**

D.2 Reviewer-specific sentiment analysis

Fine-Tuning. We fine-tune the `bert-base-uncased` model and benchmark DI-ERM against other domain generalization methods. For DI-ERM, we concatenate each review x with reviewer context m, represented by 20 randomly selected reviews from the same reviewer, using the text prompt in Figure 3.

We choose `nomic-embed-text-v1.5`, which supports a 2048-token window (compared to 512 for DistilBERT), in order to handle the long reviwer context m.

Table 8 reports the results. DI-ERM achieves the best performance, outperforming previously reported methods on the WILDS leaderboard (`https://wilds.stanford.edu/`).

```
Instruction: Classify the current review based on this
reviewer's sentiment patterns.

Current Review: [current review goes here]

Reviewer's Historical Reviews:
Review 1: [review_1] | Review 2: [review_2] | ...
```

Figure 3: Text prompt that encodes reviewer writing style

Table 8: Reviewer-specific sentiment analysis. DI-ERM (ours) achieves the highest accuracy, outperforming prior state-of-the-art.

Algorithm	**Model**	**Test Avg Acc**	**Test 10% Acc**
ERM	DistilBERT	72.0 ± 0.1	54.2 ± 0.8
GroupDRO	DistilBERT	70.0 ± 0.5	53.3 ± 0.8
CORAL	DistilBERT	71.1 ± 0.3	52.9 ± 0.8
IRM	DistilBERT	70.3 ± 0.6	52.4 ± 0.8
LISA (SOTA)	DistilBERT	70.7 ± 0.3	54.7 ± 0.0
ERM (finetune)	nomic-embed-text-v1.5	71.8 ± 0.9	54.7 ± 0.0
DI-ERM (ours, finetune)	nomic-embed-text-v1.5	73.1 ± 0.3	56.4 ± 0.8

Linear/MLP-probing. We also evaluate in a frozen-feature setting, where the language model is fixed and only a lightweight classifier is trained. Each review x is represented by its [CLS] embedding from a pretrained sentiment model (e.g., DistilBERT fine-tuned on SST-2). For reviewer context m, we average the [CLS] embeddings of all reviews written by that reviewer. The concatenated review and reviewer embeddings are then passed to a linear or a shallow MLP classifier implemented in `scikit-learn`.

Domain2Vec. Inspired by Zaheer et al. [2017], Deshmukh et al. [2018], we implement a Domain2Vec-style module to encode reviewer-specific domain information. Given a set of reviews $\{x_1, x_2, \ldots, x_n\} \sim P_{X|D=d}$ written by reviewer d, we learn a mapping

$$f(\{x_1, x_2, \ldots, x_n\}) = \rho\left(\frac{1}{n}\sum_{i=1}^{n}\phi(x_i)\right),$$

where ϕ and ρ are MLPs that map individual feature representations (extracted from pretrained model) to a latent space and then transform the aggregated feature, respectively. The resulting vector is concatenated with the review representation x to predict its sentiment label y.

Table 9 shows the result.

D.3 Image classification across styles

We evaluate our approach on the PACS benchmark, which contains four visual styles: Photo (P), Art Painting (A), Cartoon (C), and Sketch (S). To assess robustness to style variation, we test a diverse set of models from the CLIP and DINOv2 families.

For all the experiment we use the text prompt in Figure 4 as input to `DistillBERT`.

Table 10 summarizes the results. Across most domain shifts, our proposed DI-ERM method consistently outperforms standard pooling ERM, highlighting the advantage of incorporating domain-specific information into the representation.

Table 9: Sentiment classification on Amazon-WILDS with reviewer-specific signals. "Domain2Vec" denotes reviewer encoding based on a learned mean embedding. DI-ERM variants consistently outperform pooling ERM baselines.

Algorithm	Model	Test Avg Acc	Test 10% Acc
Pooling ERM (linear)	distilbert-base-uncased-finetuned-sst-2-english	67.42	48.00
DI-ERM (linear)	distilbert-base-uncased-finetuned-sst-2-english	**68.21**	48.00
Pooling ERM (MLP)	distilbert-base-uncased-finetuned-sst-2-english	67.59	48.00
DI-ERM (MLP)	distilbert-base-uncased-finetuned-sst-2-english	**68.28**	**49.33**
DI-ERM (Domain2Vec)	distilbert-base-uncased-finetuned-sst-2-english	68.21	48.00
Pooling ERM (linear)	bert-base-multilingual-uncased-sentiment	72.14	53.33
DI-ERM (linear)	bert-base-multilingual-uncased-sentiment	**73.22**	**54.67**
Pooling ERM (MLP)	bert-base-multilingual-uncased-sentiment	73.01	53.33
DI-ERM (MLP)	bert-base-multilingual-uncased-sentiment	**73.18**	**55.07**
DI-ERM (Domain2Vec)	bert-base-multilingual-uncased-sentiment	73.19	54.67

```
Domain "photo", text prompt:  "a photo"
Domain "art painting", text prompt:  "an oil painting"
Domain "cartoon", text prompt:  "a colorful cartoon"
Domain "sketch", text prompt:  "a pencil sketch"
```

Figure 4: Example of style-specific text prompts used as domain descriptions.

Notably, we observe that the performance gains from DI-ERM tend to diminish as model capacity increases. For the largest models (e.g., CLIP ViT-L/14 and DINOv2 ViT-L/14), the improvement is marginal or saturates. This trend is also observed by various empirical works, e.g. Cho et al. [2023].

Table 10: Domain generalization results on PACS using models from the CLIP and DINOv2 families. DI-ERM achieves improved accuracy over pooling ERM in most configurations, particularly for mid-sized models.

Model	Algorithm	PAC $\to$ S	ACS $\to$ P	CSP $\to$ A	SPA $\to$ C	Test Avg Acc
CLIP: vitb32	Pooling ERM (linear)	86.97	99.58	95.90	97.48	94.98
	DI-ERM (linear)	**88.06**	**99.64**	**96.29**	**97.48**	**95.37**
CLIP: vitb16	Pooling ERM (linear)	90.89	99.70	97.51	98.76	96.70
	DI-ERM (linear)	**91.09**	99.70	**97.61**	**98.76**	**96.79**
CLIP: vitl14	Pooling ERM (linear)	95.42	99.94	99.22	99.79	98.59
	DI-ERM (linear)	95.32	99.94	**99.32**	99.79	98.59
DINOv2: vits14	Pooling ERM (linear)	79.82	85.81	93.55	91.34	87.63
	DI-ERM (linear)	**80.45**	**90.00**	**94.09**	**91.60**	**89.04**
DINOv2: vitb14	Pooling ERM (linear)	87.27	95.45	97.66	94.67	93.76
	DI-ERM (linear)	**87.35**	**96.53**	**98.05**	94.50	**94.11**
DINOv2: vitl14	Pooling ERM (linear)	92.29	96.41	98.14	97.48	96.08
	DI-ERM (linear)	**92.42**	**97.37**	98.10	97.48	**96.34**

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The claims are either supported by theory statements or by reproducible experiment results.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Limitations about our practical method is described.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: Assumptions are stated in the theorem statement. Full proofs are included in the appendix.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: Important information about the experiments are in main text. Details on the experimental setup is described in the appendix.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Code is provided, common benchmark datase were used, instructions are given, the result is easily reproducible.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: See appendix and attached code.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: For many result on various benchmark, we fix the random seeds and uses sklearn, our result is deterministic and thus no error bar. For fine-tuning result, we do repeated trials, see appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: See appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The authors have read the NeurIPS Code of Ethics and confirm that this research follows the code of ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: This is a theory-oriented paper.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.

- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The paper poses no such risks

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Citations and urls are included.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve with this matter.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [No]

    Justification: LLM is used only for writing, editing and formatting purposes.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.