LOSSY REPRESENTATION COMPRESSION FOR OUT-OF-DISTRIBUTION GENERALIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Out-of-distribution (OOD) generalization aims to maintain predictive performance under distribution shifts. Despite extensive research, invariant learning methods often fail to deliver reliable improvements across diverse settings. A key limitation is that prediction invariance across training environments only encourages the learning of causal features but does not remove non-causal features from the learned representations. To address this challenge, we first show that causal and non-causal features will occupy different directions in the representation space. Encoding both produces more complex representations than encoding causal features alone. We therefore introduce a new regulariser based on the well-established lossy coding rate from information theory, leading to a novel simple and effective framework termed **Lossy Causal Feature Learning (*LoCaL*)**. Geometrically, the lossy coding rate measures the volume occupied by a representation at a given precision. Constraining this quantity encourages a lower-dimensional, more compact representation, thereby suppressing non-causal features. Extensive experiments across diverse datasets demonstrate that *LoCaL* outperforms a broad range of strong baselines, with most improvements being statistically significant.

1 INTRODUCTION

Deep learning has transformed numerous fields, including healthcare, robotics, and natural language processing. However, when deployed in real-world settings, deep learning models often fail to generalize when the test distribution differs from the distributions encountered during training. This phenomenon is commonly known as out-of-distribution (OOD) failure (Beery et al., 2018; Koh et al., 2021; Recht et al., 2019; Hendrycks et al., 2021). A canonical example is distinguishing cows from camels (Arjovsky et al., 2019; Kamath et al., 2021). In the training setting, cows are typically shown on grass and camels in deserts, a model may associate the labels with their backgrounds rather than the animals themselves. It may therefore fail when this association is reversed in the test environment (Beery et al., 2018; Xiao et al., 2021; Geirhos et al., 2020).

To improve OOD generalisation of deep models, many studies draw on insights from causality analysis, which motivate the *invariance principle* (Peters et al., 2016; Rojas-Carulla et al., 2018). This principle assumes that causal features maintain an invariant relationship with the target across different environments, whereas relationships between non-causal features and the target may vary (Heinze-Deml et al., 2018). In the cow–camel example, animal shape is a causal feature that defines the object, whereas the background which can vary across environments is non-causal. Although related work uses terms such as *spurious*, *shortcut*, *nuisance*, and *noise-related* features, their definitions vary but broadly describe closely related forms of non-causal information (Geirhos et al., 2020; Nagarajan et al., 2021; Bousmalis et al., 2016; Xiao et al., 2021).

The invariance principle has inspired influential approaches for OOD generalisation such as Invariant Risk Minimization (IRM) (Arjovsky et al., 2019) and Risk Extrapolation (REx) (Krueger et al., 2021). However, encouraging invariance across training environments does not guarantee that a model learns only causal features. Its representation may still contain non-causal features. This problem is particularly relevant to overparameterized models, which are common in modern deep learning applications: they can fit the training data perfectly without eliminating non-causal features (Zhang et al., 2017; Zhou et al., 2022). Complementing these observations, Ahuja et al. (2021) provide a clear theoretical characterization of this limitation for invariance-based OOD methods:

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

even for simple two-dimensional data with one causal feature and one non-causal feature, a linear classifier can satisfy the invariance constraint and achieve zero classification error across training environments while still relying on the non-causal feature. Learning causal features while discarding non-causal features therefore becomes a central challenge for OOD generalization.

To address this challenge, we are the first to explicitly consider the structure of learned representations. Our motivation comes from the meaningful semantic structure observed in large language models and multimodal foundation models. Recent studies provide theoretical and empirical evidence that semantic concepts can be encoded as directions in the representation spaces of large language models (Park et al., 2024; Jiang et al., 2024), with their geometric relationships reflecting categorical and hierarchical structures (Park et al., 2025). Studies of the multimodal model CLIP likewise identify interpretable concepts and show that its representations can be decomposed into semantically meaningful components (Goh et al., 2021; Gandelsman et al., 2024; Bhalla et al., 2024). Inspired by these findings, we hypothesize that causal and non-causal features may be encoded along different directions in the representation space. As shown in Table 1, when a model trained on MNIST learns only causal digit features, the representation rank of each class is approximately one. When it also learns non-causal features, the rank increases to approximately two. Although causal and non-causal features need not be perfectly separated, retaining both requires a more complex representation than retaining causal features alone. Section 4.1 provides the full experimental protocol and additional diagnostics. This observation motivates constraining representation complexity to suppress non-causal features.

However, quantifying the complexity of a finite set of samples is widely recognized as challenging (Buchanan et al., 2026). A longstanding idea in information theory is to measure complexity through *description length*: the number of bits needed to encode a given dataset (Rissanen, 1978). Although the shortest description generally has no universal closed-form expression, Ma et al. (2007) derived a tight upper bound based on *lossy coding rate* for a broad class of data, including Gaussian distributions and subspace-structured data. This measure also has an intuitive geometric interpretation: it measures the volume occupied by the data at a given coding precision. Notably, its closed-form expression makes it straightforward to compute and, as we will show, easy to incorporate into modern deep learning frameworks as a penalty on representation complexity.

We summarize the main contributions of this work as follows.

- **Conceptual contribution.** To the best of our knowledge, this work is the first to study OOD generalization by explicitly focusing on representation structure to retain causal features and discard non-causal features. Inspired by the semantic structure observed in foundation models, we hypothesize that causal and non-causal features may occupy different directions in the representation space. This perspective motivates controlling representation complexity to suppress non-causal features while preserving useful causal features.
- **Technical contribution.** We propose *Lossy Causal Feature Learning* (*LoCaL*), a simple and effective framework to jointly learn causal features and suppress non-causal features. The framework can easily incorporate diverse invariance-based objectives, including IRM and REx, without introducing additional representation branches or changing the prediction architecture. We evaluate *LoCaL* across diverse datasets, where it outperforms a broad range of strong baselines, with most improvements being statistically significant.

2 Related Work

2.1 Invariant learning and its limits

Domain generalization seeks predictors that transfer from multiple source environments to unseen targets (Zhou et al., 2023; Wang et al., 2023; Liu et al., 2023). Invariant learning builds on the idea that causal predictive relationships remain stable across environments, whereas non-causal correlations may change (Peters et al., 2016; Rojas-Carulla et al., 2018; Heinze-Deml et al., 2018). IRM seeks a representation for which the same classifier is optimal in every environment (Arjovsky et al., 2019), while REx penalizes variation in environment risks (Krueger et al., 2021). Related methods promote stability through games, gradient agreement, or gradient-variance matching (Ahuja et al., 2020; Parascandolo et al., 2021; Shi et al., 2022; Rame et al., 2022), alongside broader strategies based on meta-learning, augmentation, robust optimization, and flat minima (Li et al., 2018; Zhang

et al., 2018; Zhou et al., 2021; Sagawa et al., 2020; Cha et al., 2021). Invariant objectives nevertheless depend on environment diversity, the hypothesis class, and the optimization surrogate, and can admit solutions that fail under target shifts (Rosenfeld et al., 2021; Kamath et al., 2021; Gulrajani & Lopez-Paz, 2021; Galstyan et al., 2022). Although variants such as Bayesian IRM refine the estimation of invariance (Lin et al., 2022), stable predictions across observed environments do not determine all information retained in an overparameterized representation. *LoCaL* therefore combines invariant prediction with an explicit preference for representational compactness.

2.2 Compression, sparsity, and feature separation

The information bottleneck (IB) trades input compression against predictive information (Tishby et al., 2000; Alemi et al., 2017). IB-IRM uses compression to select among invariant predictors (Ahuja et al., 2021), while Invariant Information Bottleneck introduces a variational objective for nonlinear prediction with invariant non-causal features (Li et al., 2022). These approaches typically use variance- or distribution-based surrogates. By contrast, our lossy coding rate measures the geometry of finite learned features through their full second-moment spectrum, distinguishing representations with equal total energy but different occupied directions. Section 3.3 formalizes its relationship to variance regularization.

Other methods constrain parameter complexity or separate information sources. Modular Risk Minimization selects an invariant subnetwork, whereas SparseIRM imposes parameter sparsity (Zhang et al., 2021; Zhou et al., 2022). Architectural methods separate task and domain information through private components, disentanglement, style randomization, or specialized branches (Bousmalis et al., 2016; Zhang et al., 2022; Nam et al., 2021; Chen et al., 2024); related prompt-based methods target foundation models (Cheng et al., 2024; Son et al., 2025). In contrast, *LoCaL* neither selects a subnetwork nor assumes an explicit task–domain decomposition: it directly regularizes the encoder's predictive representation and applies to invariant objectives such as IRM and REx.

2.3 Representation geometry and lossy coding

Evidence from language models and CLIP suggests that learned concepts can occupy meaningful directions or semantic components in representation space (Park et al., 2024; Jiang et al., 2024; Park et al., 2025; Goh et al., 2021; Gandelsman et al., 2024; Bhalla et al., 2024). We do not assume perfect disentanglement or coordinate-wise interpretability. We use only the weaker hypothesis tested in Section 3.2: retaining causal and non-causal information can require additional representation directions.

Lossy coding quantifies this structure at finite precision (Ma et al., 2007), connecting minimum description length (Rissanen, 1978) to the volume and spectrum of sample representations. Coding-rate objectives have been used to learn diverse, discriminative representations through maximal coding-rate reduction (Yu et al., 2020; Chan et al., 2022; Baek et al., 2022; Yu et al., 2024). In domain generalization, Localized Adversarial Domain Generalization uses coding rate to counter feature collapse during adversarial alignment (Zhu et al., 2022). *LoCaL* instead minimizes the coding cost of the pooled predictive representation subject to prediction and invariance objectives, yielding a spectrum-sensitive representation bottleneck.

3 Lossy Causal Feature Learning

3.1 Learning Guided by the Invariance Principle

The invariance principle seeks predictive relationships that remain stable across environments (Peters et al., 2016; Rojas-Carulla et al., 2018). Methods based on this principle typically combine prediction loss with a penalty that encourages consistent behavior across training environments (Arjovsky et al., 2019; Krueger et al., 2021). We use such objectives as the foundation of *LoCaL*.

Let $\mathcal{E}_{\text{tr}}$ denote the set of training environments. An encoder $f_{\boldsymbol{\Theta}}$ maps an input $\boldsymbol{x}$ to a representation $\boldsymbol{z} \in \mathbb{R}^d$, and a classifier $g_{\boldsymbol{w}}$ maps the representation to a prediction: $f_{\boldsymbol{\Theta}}(\boldsymbol{x}) = \boldsymbol{z}$, $g_{\boldsymbol{w}}(\boldsymbol{z}) = \hat{y}$. The empirical risk in environment e is $\widehat{R}_e(\boldsymbol{\Theta}, \boldsymbol{w}) = \frac{1}{n_e} \sum_{i=1}^{n_e} \ell\left(g_{\boldsymbol{w}}\left(f_{\boldsymbol{\Theta}}(\boldsymbol{x}_i^e)\right), y_i^e\right)$.

IRMv1 (Arjovsky et al., 2019) encourages the same classifier to be optimal across environments:

$$\mathcal{L}_{\text{IRMv1}} = \sum_{e \in \mathcal{E}_{\text{tr}}} \widehat{R}_e(\boldsymbol{\Theta}, \boldsymbol{w}) + \beta \sum_{e \in \mathcal{E}_{\text{tr}}} \left| \nabla_\gamma \widehat{R}_e^\gamma(\boldsymbol{\Theta}, \boldsymbol{w}) \Big|_{\gamma=1} \right|^2, \tag{1}$$

where $\widehat{R}_e^\gamma$ denotes the environment risk obtained by scaling the classifier logits by a dummy scalar γ. The gradient penalty tests whether the same reference scale, $\gamma = 1$, is locally optimal across all training environments. REx (Krueger et al., 2021) instead penalizes variation in prediction risk across environments:

$$\mathcal{L}_{\text{REx}} = \sum_{e \in \mathcal{E}_{\text{tr}}} \widehat{R}_e(\boldsymbol{\Theta}, \boldsymbol{w}) + \beta \, \text{Var}_{e \in \mathcal{E}_{\text{tr}}} \left[\widehat{R}_e(\boldsymbol{\Theta}, \boldsymbol{w}) \right]. \tag{2}$$

We use $\mathcal{L}_{\text{inv}}$ to denote either objective. Both encourage invariant prediction, but neither explicitly requires the learned representation to discard non-causal features. We next use a controlled example to illustrate this limitation of invariance-based methods.

3.2 A MOTIVATING EXAMPLE

We present a motivating experiment to establish two points. First, invariant learning can retain non-causal features even when causal features are sufficient for prediction. Second, encoding those additional features can increase representation complexity. Section 4.1 describes the experimental details and Appendix A provides complete implementation details.

We train two matched IRMv1 models on binary MNIST with two environments. Both tasks use black digits on randomly sampled red or green backgrounds, so background color is independent of digit identity and environment. The first task uses the correct labels. In the second, 25% of the training labels are independently flipped. Because digit identity alone cannot predict which labels were corrupted, fitting substantially more than 75% of the noisy training labels requires the model to encode additional, non-causal information.

Table 1: Performance and representation complexity for the matched IRMv1 controls and *LoCaL*.

	Performance (%)		Rank on Env 1		Rank on Env 2		Rank on Env 1&2	
Method	Train Acc.	Test Acc.	Digit 0	Digit 1	Digit 0	Digit 1	Digit 0	Digit 1
Correct labels (IRMv1)	100.000	99.903	1	1	1	1	1	1
25% label flips (IRMv1)	99.906	66.764	2	2	2	2	2	2
25% label flips (*LoCaL*)	90.264	**78.443**	1	1	1	1	1	1

Table 1 shows that the noisy-label model attains 99.906% training accuracy, above the 75% achievable from digit identity alone, but only 66.764% accuracy on gray-background test images. It therefore retains the causal digit feature while exploiting additional non-causal information. Correspondingly, the class-specific 75% energy rank increases from 1 to 2 for both digits in each environment and after pooling. This threshold provides a clear comparison rather than an exact characterization of rank. Appendix A reports the full spectra and computation details. Besides, Figure 1 compares normalized SmoothGrad saliency maps computed with 64 shared perturbations across the two IRMv1 controls. The correct-label IRMv1 model responds more locally around the digit but the noisy-label model more diffusely to the background. Together, these results show that predictive invariance need not remove non-causal information and that retaining such information can expand the representation subspace.

3.3 FROM REPRESENTATION DIMENSION TO LOSSY CODING RATE

The above example shows that a model can retain non-causal predictive information alongside causal features, producing a more complex representation. The key challenge is finding a tractable measure of the information encoded in the representation. This idea builds on the information-theoretic principle of *minimum description length*, which measures data complexity by the number of bits required to encode it (Rissanen, 1978). Although no universal closed-form description length exists for arbitrary data, Ma et al. (2007) derived a closed-form *lossy coding rate* measure for finite samples

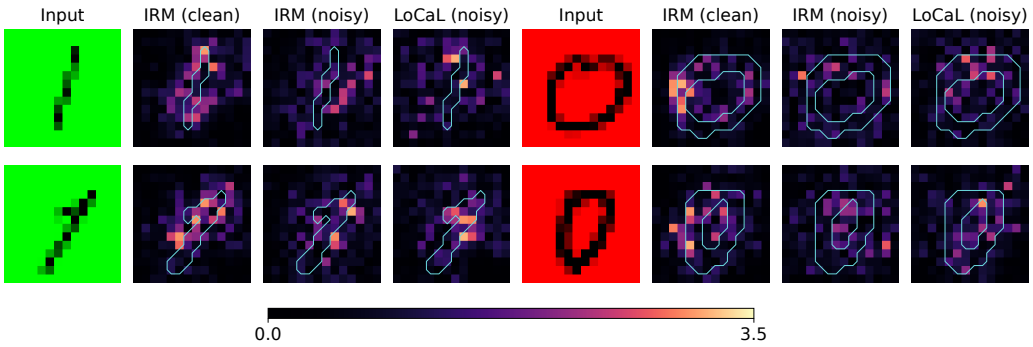

Figure 1: SmoothGrad maps for four training images with green (left) and red (right) backgrounds.

from broad distribution families, including Gaussian and subspace-structured data. For a batch representation matrix $\boldsymbol{Z} \in \mathbb{R}^{d\times m}$, the lossy coding rate is

$$\mathcal{R}_\epsilon(\boldsymbol{Z}) = \frac{1}{2}\log\det\left(\boldsymbol{I}_d + \frac{d}{m\epsilon^2}\boldsymbol{Z}\boldsymbol{Z}^\top\right), \tag{3}$$

where $\epsilon > 0$ specifies the coding precision (Ma et al., 2007).

This quantity has two complementary interpretations. From an information-theoretic perspective, with a base-2 logarithm, it quantifies the number of bits needed to encode $\boldsymbol{Z}$ at the specified precision. In practice, implementations commonly use the natural logarithm for convenience. From a geometric perspective, it measures the volume occupied by $\boldsymbol{Z}$ in the representation space. Minimizing the lossy coding rate $\mathcal{R}_\epsilon(\boldsymbol{Z})$ encourages $\boldsymbol{Z}$ to occupy a smaller volume, thereby yielding a more compact representation.

The geometric view also clarifies the relationship between lossy coding rate and representation rank. Let σ_i denote the singular values of $\boldsymbol{Z}$ and let $c = d/(m\epsilon^2)$. Then

$$\mathcal{R}_\epsilon(\boldsymbol{Z}) = \frac{1}{2}\sum_i \log(1 + c\sigma_i^2) \geq \frac{1}{2}\log\left(1 + c\sum_i \sigma_i^2\right). \tag{4}$$

The equality follows from the singular-value decomposition of $\boldsymbol{Z}$, while the inequality follows from $\prod_i(1 + c\sigma_i^2) \geq 1 + c\sum_i \sigma_i^2$. For fixed nonzero feature energy $\sum_i \sigma_i^2$, equality is attained when all energy is concentrated in a single singular direction. Distributing the same energy across additional directions increases the coding rate. Thus, when feature scale is controlled by the predictive objective and training procedure, minimizing the coding rate encourages a more concentrated spectrum and a lower effective rank.

3.4 The *LoCaL* objective

We now show how can this lossy coding rate be incorporated into existing invariant learning objectives. The arguably most natural way is to constrain the information encoded within each class through lossy coding rate. Let $\boldsymbol{Z}_k$ contain the m_k representations from class k, let $m = \sum_k m_k$, and define $p_k = m_k/m$. The corresponding class-wise objective is

$$\min_{\boldsymbol{\Theta},\boldsymbol{w}} \quad \mathcal{L}_{\text{inv}} + \lambda\mathcal{R}_{\epsilon,c}(\boldsymbol{Z}), \qquad \mathcal{R}_{\epsilon,c}(\boldsymbol{Z}) = \sum_k p_k\mathcal{R}_\epsilon(\boldsymbol{Z}_k). \tag{5}$$

It is clear that the invariance objective $\mathcal{L}_{\text{inv}}$ promotes learning causal features, while the class-wise coding penalty $\mathcal{R}_{\epsilon,c}(\boldsymbol{Z})$ limits non-causal features retained in the representation.

However, computing Equation (5) requires a separate log-determinant for each class, which can be expensive when the number of classes is large. Fortunately, at a common coding precision, the class-wise penalty has a tight upper bound based on the pooled representation $\boldsymbol{Z}$:

$$\mathcal{R}_{\epsilon,c}(\boldsymbol{Z}) = \sum_k p_k\mathcal{R}_\epsilon(\boldsymbol{Z}_k) \leq \mathcal{R}_\epsilon(\boldsymbol{Z}). \tag{6}$$

Equality holds when all nonempty classes share the same second-moment matrix, $\boldsymbol{Z}_k \boldsymbol{Z}_k^\top / m_k$. Using this upper bound replaces the class-wise log-determinants with a single calculation and yields the final *Lossy Causal Feature Learning* (*LoCaL*) objective:

$$\min_{\boldsymbol{\Theta}, \boldsymbol{w}} \quad \mathcal{L}_{\text{inv}} + \lambda \mathcal{R}_\epsilon(\boldsymbol{Z}), \tag{7}$$

where $\lambda \geq 0$ controls the strength of representation compression.

Relationship to the information bottleneck regulariser. Ahuja et al. (2021) theoretically show that the invariance principle can be insufficient even in a simple two-dimensional classification problem with one causal feature and one non-causal feature. They therefore introduce an information-bottleneck (IB) constraint to select, among predictive invariant representations, one that retains less information. Because representation entropy is difficult to optimize, its practical objective replaces entropy with feature variance, using the maximum-entropy property of the Gaussian distribution to justify variance as an upper-bound surrogate.

This perspective asks how much information a representation retains, but it does not explicitly characterize how that information is organized within a given representation space. In particular, a variance penalty reduces the representation to a single scalar and does not distinguish the number of occupied directions, their relative strengths, or the resulting effective rank. The theoretical analysis in (Ahuja et al., 2021) focuses on linear models, while its variance surrogate in higher-dimensional neural representations is supported empirically. By contrast, *LoCaL* makes representation structure the central object of study: its coding-rate penalty depends on the full spectrum of the representation and therefore captures both the amount of variation and how that variation is distributed across directions. To the best of our knowledge, this is the first OOD framework to explicitly control representation structure for retaining causal features while suppressing non-causal features.

The following result makes the relationship between the two penalties precise.

Proposition 3.1 (Variance as a first-order approximation)**.** *Let $\boldsymbol{M_Z} = \boldsymbol{Z}\boldsymbol{Z}^\top / m$ have eigenvalues μ_j, and let $\rho = d/\epsilon^2$. Then*

$$\mathcal{R}_\epsilon(\boldsymbol{Z}) = \frac{1}{2} \sum_j \log(1 + \rho \mu_j). \tag{8}$$

When $\rho \max_j \mu_j \to 0$, or equivalently when ϵ^2 is significantly large relative to the feature spectrum,

$$\mathcal{R}_\epsilon(\boldsymbol{Z}) = \frac{\rho}{2} \operatorname{tr}(\boldsymbol{M_Z}) - \frac{\rho^2}{4} \operatorname{tr}(\boldsymbol{M_Z^2}) + O\left(\rho^3 \sum_j \mu_j^3\right). \tag{9}$$

Proof. The eigenvalues of $\boldsymbol{M_Z}$ are μ_j, which gives Equation (8). Applying $\log(1 + x) = x - x^2/2 + O(x^3)$ to each $\rho\mu_j$ gives Equation (9). □

For centred representations, the leading term $\operatorname{tr}(\boldsymbol{M_Z})$ is exactly the total feature variance. The variance penalty is therefore only a first-order approximation to the *LoCaL* coding penalty, omitting higher-order spectral terms that distinguish representations with the same total variance but different directions and effective ranks. Outside the asymptotic regime $\rho \max_j \mu_j \to 0$, this approximation can be loose. In our numerical analysis with a centered random representation $\boldsymbol{Z} \in \mathbb{R}^{128 \times 1024}$, reducing the relative approximation error below 1% requires $\epsilon^2 \approx 7{,}172$, which is on the order of 10^4. In practice, however, the coding precision is typically much smaller (Buchanan et al., 2026); we fix $\epsilon^2 = 10$ in all experiments, as reported in Appendix B. Thus, variance can be a poor surrogate for the coding penalty at practical precision levels, which may help explain the weaker and less stable performance of IB-IRM relative to *LoCaL* in our experiments. Appendix C.3 provides the complete numerical analysis.

4 Experiments

Experimental setup. We consider two settings introduced in (Ahuja et al., 2021). In the *partially informative invariant feature* (PIIF) setting, invariant features determine only part of the label, allowing overparameterized models to fit the remaining training examples using non-causal features.

Full-AC-CMNIST (Zhou et al., 2022), CIFAR-MNIST (Krueger et al., 2021), and Colored Object (Gulrajani & Lopez-Paz, 2021) belong to this setting. In the *fully informative invariant feature* (FIIF) setting, invariant features determine the complete label, although a model may still retain predictive non-causal features. CS-MNIST (Ahuja et al., 2021) represents this setting.

Dataset. Full-AC-CMNIST extends anti-causal Colored MNIST (Arjovsky et al., 2019) to ten classes. Labels are flipped with probability 0.2, and color agrees with the resulting label with probabilities $(0.999, 0.7, 0.1)$ in the two training environments and the test environment, respectively. CIFAR-MNIST concatenates a CIFAR-10 image, which provides the invariant component, with an MNIST image, which provides the non-causal component. It uses 10% label noise and the correlation tuple $(0.999, 0.8, 0.1)$. Colored Object places objects from eight MSCOCO classes on colored backgrounds using the same noise level and correlation tuple. CS-MNIST assigns a distinct color to each digit, uses two training environments containing $20{,}000$ examples with correlations $(1.0, 0.9)$, and evaluates on $20{,}000$ randomly colored test examples.

Baselines. We compare *LoCaL* against empirical risk minimization (ERM), IRMv1 (Arjovsky et al., 2019), REx (Krueger et al., 2021), Bayesian IRM (Lin et al., 2022), modular risk minimization (MRM) (Zhang et al., 2021), SparseIRM (Zhou et al., 2022), and the variance-based information bottleneck (IB) (Ahuja et al., 2021). MRM and SparseIRM constrain subnetworks of an overparameterized model, whereas IB and *LoCaL* regularize the learned representation. Oracle denotes the corresponding benchmark without non-causal correlations and is included only as a reference.

In the following performance tables, boldface indicates the highest mean within each method family for each dataset. "Gain" is calculated as the larger of the two *LoCaL* means minus the highest mean among all non-Oracle, non-*LoCaL* baselines. The reported p-values are obtained from two-sided paired t-tests.

4.1 Analysis of non-causal feature retention

We use the controlled binary-MNIST experiment from Section 3.2 to evaluate whether *LoCaL* can suppress the non-causal information retained by invariant learning. We retain the $10{,}610$ digit-0 and digit-1 images from the first $50{,}000$ MNIST training examples and divide them equally between two environments. Each digit is rendered in black on a red or green background sampled independently of both digit identity and environment, and 25% of the training labels are independently flipped. We apply *LoCaL* to IRMv1 using the same lightweight CNN and optimization schedule as the matched IRMv1 controls. The 128-dimensional penultimate activations define the representation, and we directly set $\epsilon^2 = 10$ and $\lambda = 100$.

The held-out set contains $2{,}055$ correctly labelled digits on gray backgrounds. We evaluate the class-specific spectra of the raw, uncentered training representations using the 75% energy rank in Table 1, and compare the SmoothGrad maps in Figure 1. Appendix A provides the complete data construction, optimization protocol, and singular-value spectra.

Section 3.2 shows that noisy-label IRMv1 retains non-causal information in addition to the causal digit feature. Here, we focus on the effect of adding the coding-rate penalty. *LoCaL* improves gray-background test accuracy from 66.764% to 78.443%, while reducing every reported class-specific energy rank from 2 to 1. Its SmoothGrad maps are also more concentrated along the digit contours and less diffuse over the background than those of noisy-label IRMv1. Together, these results suggest that constraining representation complexity effectively suppresses non-causal background information and redirects the model toward the causal digit shape.

4.2 Robustness to increasing overparameterization

We first follow the Full-AC-CMNIST setting of Zhou et al. (2022). A three-layer multilayer perceptron (MLP) uses a common hidden width ranging from 64 to 1024. We consider training sets of $1{,}500$ and $1{,}000$ examples, denoted Full-AC-CMNIST-1500 and Full-AC-CMNIST-1000, respectively. Even at width 64, both settings are substantially overparameterized. Regularization coefficients are selected using a 10% split of the training data, after which test performance is evaluated over five runs. Appendix B.2 provides the complete optimization and selection details.

Table 2: Test accuracy (%) on Full-AC-CMNIST-1500 across MLP hidden widths.

Hidden width		64	128	256	384	512	768	896	1024
Oracle		75.03 ± 0.52	75.55 ± 0.48	75.79 ± 0.54	75.94 ± 0.54	75.98 ± 0.55	76.06 ± 0.50	76.15 ± 0.49	76.22 ± 0.63
ERM		31.64 ± 0.45	32.54 ± 0.20	33.25 ± 0.42	34.48 ± 0.53	35.73 ± 0.53	37.02 ± 0.52	37.32 ± 0.61	38.23 ± 0.85
Bayesian IRM		53.59 ± 1.43	59.22 ± 1.37	59.14 ± 0.62	56.67 ± 1.00	53.38 ± 1.51	47.77 ± 1.47	47.31 ± 1.35	41.44 ± 1.26
IRMv1	Plain	56.19 ± 1.23	59.44 ± 0.89	58.50 ± 0.56	57.12 ± 0.88	55.49 ± 1.65	51.22 ± 1.01	49.45 ± 1.40	47.47 ± 0.52
	MRM	56.97 ± 0.73	58.76 ± 1.00	58.32 ± 1.03	57.03 ± 0.74	55.18 ± 0.97	51.31 ± 0.64	49.65 ± 0.80	47.63 ± 0.86
	IB	55.09 ± 1.54	59.16 ± 0.98	59.08 ± 0.80	57.87 ± 0.94	**58.81** ± 0.39	54.20 ± 2.20	56.83 ± 2.01	53.72 ± 1.85
	Sparse	**57.63** ± 0.68	58.98 ± 0.85	57.94 ± 1.29	57.86 ± 1.60	58.37 ± 1.54	58.09 ± 1.46	58.00 ± 0.95	57.77 ± 0.96
	LoCaL	56.17 ± 1.23	**59.45** ± 0.88	**60.94** ± 0.66	**58.41** ± 1.31	57.33 ± 1.48	**58.47** ± 0.87	**58.80** ± 0.30	**58.34** ± 1.16
REx	Plain	60.34 ± 1.24	62.84 ± 0.80	62.37 ± 0.52	60.22 ± 0.77	57.98 ± 1.15	54.27 ± 1.20	51.95 ± 1.35	50.82 ± 1.29
	MRM	60.65 ± 0.88	62.94 ± 0.99	62.31 ± 0.84	60.21 ± 0.75	58.49 ± 1.07	56.64 ± 0.70	55.79 ± 0.80	55.15 ± 1.05
	IB	60.11 ± 0.77	**64.44** ± 0.84	64.37 ± 0.63	62.40 ± 0.39	60.87 ± 0.46	54.70 ± 1.00	51.50 ± 0.95	49.47 ± 0.54
	Sparse	61.15 ± 0.87	62.38 ± 0.92	61.81 ± 0.83	60.72 ± 1.45	60.14 ± 0.82	58.20 ± 1.34	57.38 ± 0.64	56.55 ± 0.42
	LoCaL	**61.80** ± 0.89	63.52 ± 0.94	**64.92** ± 0.71	**64.46** ± 1.00	**62.32** ± 1.01	**65.84** ± 0.63	**65.19** ± 0.79	**63.55** ± 0.55
Gain		↑ 0.65	↓ 0.92	↑ 0.55	↑ 2.06	↑ 1.45	↑ 7.64	↑ 7.19	↑ 5.78
p-value		0.4429	—	0.1843	0.0120	0.0352	< 0.0001	< 0.0001	< 0.0001

Table 3: Test accuracy (%) on Full-AC-CMNIST-1000 across MLP hidden widths.

Hidden width		64	128	256	384	512	768	896	1024
Oracle		74.42 ± 0.48	74.84 ± 0.27	75.27 ± 0.32	75.29 ± 0.25	75.44 ± 0.30	75.53 ± 0.32	75.57 ± 0.29	75.64 ± 0.28
ERM		32.25 ± 1.04	32.76 ± 0.56	33.55 ± 0.82	34.46 ± 1.13	35.64 ± 1.22	36.98 ± 0.68	37.69 ± 0.72	38.02 ± 0.49
Bayesian IRM		51.69 ± 1.78	56.86 ± 0.65	55.03 ± 1.40	50.28 ± 0.77	45.21 ± 1.40	41.46 ± 2.40	40.67 ± 2.95	41.00 ± 1.45
IRMv1	Plain	54.11 ± 1.02	56.30 ± 0.94	54.55 ± 0.56	52.18 ± 1.07	49.06 ± 1.01	43.86 ± 1.50	41.62 ± 1.51	40.81 ± 0.62
	MRM	53.40 ± 1.34	56.31 ± 0.66	55.06 ± 0.37	52.60 ± 0.98	49.38 ± 1.20	44.02 ± 1.16	42.26 ± 1.05	40.90 ± 1.52
	IB	53.16 ± 1.09	55.87 ± 0.79	**55.69** ± 0.42	53.88 ± 0.92	**55.62** ± 2.51	**55.85** ± 1.31	54.92 ± 0.37	52.77 ± 0.77
	Sparse	**54.46** ± 1.13	**56.54** ± 0.69	53.33 ± 2.38	53.07 ± 1.23	55.20 ± 1.22	53.99 ± 0.90	53.70 ± 0.43	53.58 ± 0.45
	LoCaL	54.08 ± 1.03	56.44 ± 0.79	55.06 ± 0.94	**53.91** ± 1.20	54.85 ± 1.32	55.34 ± 1.27	**55.23** ± 1.29	**54.37** ± 1.28
REx	Plain	59.16 ± 0.81	60.14 ± 0.72	56.87 ± 0.66	53.45 ± 0.99	49.94 ± 0.28	45.80 ± 0.70	46.06 ± 0.67	46.72 ± 0.39
	MRM	59.39 ± 1.25	59.73 ± 0.61	57.35 ± 0.58	54.72 ± 1.08	53.88 ± 1.14	51.25 ± 0.98	50.48 ± 1.39	50.00 ± 1.26
	IB	58.29 ± 1.51	**61.54** ± 0.64	59.23 ± 0.67	56.72 ± 1.32	54.61 ± 0.78	46.10 ± 1.14	45.62 ± 0.54	46.33 ± 0.38
	Sparse	59.49 ± 0.56	58.79 ± 0.80	56.61 ± 0.88	54.58 ± 1.01	54.00 ± 0.94	51.73 ± 1.52	50.46 ± 0.92	49.62 ± 0.58
	LoCaL	**59.65** ± 0.85	60.84 ± 0.73	**61.37** ± 0.43	**58.99** ± 1.07	**55.37** ± 1.21	**62.34** ± 1.11	**59.85** ± 0.53	**56.99** ± 0.47
Gain		↑ 0.16	↓ 0.70	↑ 2.14	↑ 2.27	↓ 0.25	↑ 6.49	↑ 4.93	↑ 3.41
p-value		0.7146	—	0.0017	0.0971	—	0.0012	< 0.0001	0.0003

Table 4: Test accuracy (%) on diverse dataset.

Method		CIFAR-MNIST	Colored Object	CS-MNIST
Oracle		83.20 ± 1.07	83.81 ± 0.72	85.65 ± 0.11
ERM		15.50 ± 0.97	35.24 ± 1.09	59.11 ± 1.14
Bayesian IRM		56.22 ± 1.23	70.09 ± 2.54	58.45 ± 1.29
IRMv1	Plain	52.56 ± 2.06	53.02 ± 4.69	59.77 ± 1.16
	MRM	53.62 ± 3.20	63.57 ± 2.76	56.28 ± 1.93
	Sparse	56.30 ± 2.34	64.32 ± 4.06	60.00 ± 4.05
	IB	55.32 ± 2.84	66.52 ± 2.96	70.80 ± 1.02
	LoCaL	**58.50** ± 2.18	**68.13** ± 2.65	**71.12** ± 1.09
REx	Plain	54.08 ± 2.86	70.19 ± 3.58	63.94 ± 1.10
	MRM	53.30 ± 1.62	67.82 ± 1.24	61.22 ± 1.93
	Sparse	54.46 ± 1.13	68.41 ± 3.82	67.95 ± 3.78
	IB	55.32 ± 2.84	70.02 ± 2.96	71.16 ± 1.03
	LoCaL	**58.56** ± 3.24	**72.78** ± 2.37	**74.26** ± 0.98
Gain		↑ 2.26	↑ 2.59	↑ 3.10
p-value		0.0778	0.0469	0.0003

Table 2 shows that the plain invariant objectives become less reliable as model width increases. Plain REx peaks at width 128 and declines to 50.82% at width 1024, whereas REx–*LoCaL* retains 63.55% accuracy. REx–*LoCaL* obtains the highest mean within the REx block at seven of the eight widths, while IRMv1–*LoCaL* leads its block at six widths. Relative to the strongest non-*LoCaL* baseline, the gains reach 7.64, 7.19, and 5.78 percentage points at widths 768, 896, and 1024, respectively, and are statistically significant at each of these widths ($p < 0.0001$). These results indicate that representation compression is particularly beneficial in the most overparameterized models. Table 3 considers the more data-limited Full-AC-CMNIST-1000 setting. REx–*LoCaL* again achieves the highest mean within the REx block at seven of the eight widths; the only exception is width 128. It is also the strongest non-Oracle method overall at six widths. At widths 768, 896, and 1024, it obtains 62.34%, 59.85%, and 56.99%, exceeding the strongest non-*LoCaL* baselines by 6.49, 4.93, and 3.41 percentage points, respectively; all three gains are statistically significant

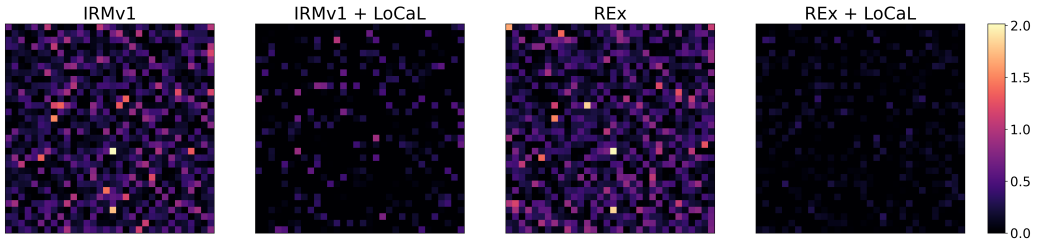

Figure 2: Feature responses to a color intervention on one Full-AC-CMNIST-1500 test image. Each 1024-dimensional absolute difference vector is reshaped into a 32×32 array for display.

($p \leq 0.0012$). The IRMv1 results are less consistent, with IRMv1–*LoCaL* attaining the highest mean within its block at three widths. Overall, the results continue to favor combining *LoCaL* with REx in the higher-dimensional regime.

4.3 Evaluation across datasets

CIFAR-MNIST and Colored Object use ResNet-18, whereas CS-MNIST uses a three-layer convolutional network. All regularization coefficients are selected using training-derived validation splits; architecture, optimization, and hyperparameter details are given in Appendix B. As shown in Table 4, *LoCaL* achieves the highest mean within both the IRMv1 and REx blocks on all three datasets. The best *LoCaL* variant improves over the strongest non-*LoCaL* baseline by 2.26, 2.59, and 3.10 percentage points on CIFAR-MNIST, Colored Object, and CS-MNIST, respectively. The gains are statistically significant on Colored Object ($p = 0.0469$) and CS-MNIST ($p = 0.0003$), while the CIFAR-MNIST improvement does not reach the 0.05 level ($p = 0.0778$). The consistent positive gains across PIIF and FIIF settings suggest that representation compression complements predictive invariance under different forms of non-causal correlation.

4.4 Feature response to color perturbations

To examine sensitivity to a non-causal intervention, we select one digit-0 test image and alter only its color. For each width-1024 Full-AC-CMNIST-1500 model, we compute the absolute coordinate-wise difference between the original and perturbed representations and reshape the resulting vector into a 32×32 display. The IRMv1 and REx representations respond more strongly to the color intervention than their corresponding *LoCaL* variants. This is consistent with *LoCaL* suppressing non-causal color information. Because the analysis uses only one example and the reshape imposes no spatial topology on the feature coordinates, the result should be interpreted as a qualitative diagnostic rather than a population-level estimate.

5 Conclusion

This work revisits a limitation of invariant learning: stable predictions across observed environments do not ensure that non-causal information is removed from the learned representation. Our motivating experiment shows that retaining both causal and non-causal features increases representation complexity. Based on this observation, we introduced *Lossy Causal Feature Learning* (*LoCaL*), which augments an invariant objective with a lossy-coding-rate penalty. By operating on the full representation spectrum, *LoCaL* encourages information to concentrate in fewer directions and can be combined directly with invariant learners such as IRM and REx.

Across partially and fully informative invariant-feature settings, *LoCaL* improves OOD performance over invariance, sparsity, and variance-based regularization methods, with particularly consistent gains when combined with REx and as model width increases. Representation diagnostics are also consistent with reduced sensitivity to non-causal color information. These findings support representational compactness as a complement to predictive invariance. Future work should examine the dependence of coding rate on feature scale and second-order statistics in larger models and more complex environments.

Reproducibility Statement

The experimental settings, model architectures, optimization procedures, hyperparameters, and evaluation metrics are documented in the paper and appendix. All experiments use publicly available datasets, and the reported results are aggregated over multiple random seeds where applicable. The code is in supplementary material.

Ethics Statement

This work studies general methods for out-of-distribution generalization and does not involve human participants, personally identifiable information, or private data. All experiments are conducted using publicly available benchmark datasets. Although improved robustness may benefit the reliability of machine-learning systems, our method does not eliminate all forms of distributional bias or guarantee robustness under arbitrary shifts. Models should therefore be evaluated carefully before deployment in safety-critical or socially sensitive applications.

AI Use Statement

The authors used OpenAI GPT-5.6 Sol and GPT-6 Astra to assist with language polishing, LaTeX formatting, and writing experimental code. The authors reviewed and validated all AI-assisted content and remain fully responsible for the scientific claims, experimental results, and final manuscript.

References

Kartik Ahuja, Karthikeyan Shanmugam, Kush Varshney, and Amit Dhurandhar. Invariant risk minimization games. In *International Conference on Machine Learning*, pp. 145–155, 2020. URL https://proceedings.mlr.press/v119/ahuja20a.html.

Kartik Ahuja, Ethan Caballero, Dinghuai Zhang, Jean-Christophe Gagnon-Audet, Yoshua Bengio, Ioannis Mitliagkas, and Irina Rish. Invariance principle meets information bottleneck for out-of-distribution generalization. In *Advances in Neural Information Processing Systems*, volume 34, pp. 3438–3450, 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/hash/1c336b8080f82bcc2cd2499b4c57261d-Abstract.html.

Alexander A. Alemi, Ian Fischer, Joshua V. Dillon, and Kevin Murphy. Deep variational information bottleneck. In *International Conference on Learning Representations*, 2017. URL https://research.google/pubs/deep-variational-information-bottleneck/.

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019. URL https://arxiv.org/abs/1907.02893.

Christina Baek, Ziyang Wu, Kwan Ho Ryan Chan, Tianjiao Ding, Yi Ma, and Benjamin D Haeffele. Efficient maximal coding rate reduction by variational forms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 500–508, 2022.

Sara Beery, Grant Van Horn, and Pietro Perona. Recognition in terra incognita. In *European Conference on Computer Vision*, 2018. URL https://arxiv.org/abs/1807.04975.

Usha Bhalla, Alex Oesterling, Suraj Srinivas, Flavio P. Calmon, and Himabindu Lakkaraju. Interpreting CLIP with sparse linear concept embeddings (SpLiCE). In *Advances in Neural Information Processing Systems*, volume 37, pp. 84298–84328, 2024. doi: 10.52202/079017-2678. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/996bef37d8a638f37bdfcac2789e835d-Abstract-Conference.html.

Konstantinos Bousmalis, George Trigeorgis, Nathan Silberman, Dilip Krishnan, and Dumitru Erhan. Domain separation networks. In *Advances in Neural Information Processing Systems*, volume 29, 2016. URL https://papers.neurips.cc/paper_files/paper/2016/hash/45fbc6d3e05ebd93369ce542e8f2322d-Abstract.html.

Sam Buchanan, Druv Pai, Peng Wang, and Yi Ma. Principles and practice of deep representation learning: or a mathematical theory of memory. *arXiv preprint arXiv:2606.06624*, 2026.

Junbum Cha, Sanghyuk Chun, Kyungjae Lee, Han-Cheol Cho, Seunghyun Park, Yunsung Lee, and Sungrae Park. SWAD: Domain generalization by seeking flat minima. In *Advances in Neural Information Processing Systems*, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/bcb41ccdc4363c6848a1d760f26c28a0-Abstract.html.

Kwan Ho Ryan Chan, Yaodong Yu, Chong You, Haozhi Qi, John Wright, and Yi Ma. Redunet: A white-box deep network from the principle of maximizing rate reduction. *Journal of machine learning research*, 23(114):1–103, 2022.

Liang Chen, Yong Zhang, Yibing Song, Zhen Zhang, and Lingqiao Liu. A causal inspired early-branching structure for domain generalization. *International Journal of Computer Vision*, 132(9): 4052–4072, 2024. URL https://arxiv.org/abs/2403.08649.

De Cheng, Zhipeng Xu, Xinyang Jiang, Nannan Wang, Dongsheng Li, and Xinbo Gao. Disentangled prompt representation for domain generalization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23595–23604, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Cheng_Disentangled_Prompt_Representation_for_Domain_Generalization_CVPR_2024_paper.html.

Tigran Galstyan, Hrayr Harutyunyan, Hrant Khachatrian, Greg Ver Steeg, and Aram Galstyan. Failure modes of domain generalization algorithms. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. URL https://arxiv.org/abs/2111.13733.

Yossi Gandelsman, Alexei A. Efros, and Jacob Steinhardt. Interpreting CLIP's image representation via text-based decomposition. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/5085a2b8c298edeadc46b9ffe6df5f64-Abstract-Conference.html.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2:665–673, 2020. doi: 10.1038/s42256-020-00257-z. URL https://arxiv.org/abs/2004.07780.

Gabriel Goh, Nick Cammarata, Chelsea Voss, Shan Carter, Michael Petrov, Ludwig Schubert, Alec Radford, and Chris Olah. Multimodal neurons in artificial neural networks. *Distill*, 2021. doi: 10.23915/distill.00030. URL https://distill.pub/2021/multimodal-neurons/.

Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. In *International Conference on Learning Representations*, 2021. URL https://arxiv.org/abs/2007.01434.

Christina Heinze-Deml, Jonas Peters, and Nicolai Meinshausen. Invariant causal prediction for nonlinear models. *Journal of Causal Inference*, 6(2), 2018. doi: 10.1515/jci-2017-0016. URL https://arxiv.org/abs/1706.08576.

Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *IEEE/CVF International Conference on Computer Vision*, pp. 8340–8349, 2021. URL https://openaccess.thecvf.com/content/ICCV2021/html/Hendrycks_The_Many_Faces_of_Robustness_A_Critical_Analysis_of_Out-of-Distribution_ICCV_2021_paper.html.

Yibo Jiang, Goutham Rajendran, Pradeep Kumar Ravikumar, Bryon Aragam, and Victor Veitch. On the origins of linear representations in large language models. In *International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 21879–21911, 2024. URL https://proceedings.mlr.press/v235/jiang24d.html.

Pritish Kamath, Akilesh Tangella, Danica Sutherland, and Nathan Srebro. Does invariant risk minimization capture invariance? In *International Conference on Artificial Intelligence and Statistics*, pp. 4069–4077, 2021. URL https://proceedings.mlr.press/v130/kamath21a.html.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, Etienne David, Ian Stavness, Wei Guo, Berton Earnshaw, Imran Haque, Sara M. Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. WILDS: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*, pp. 5637–5664, 2021. URL https://proceedings.mlr.press/v139/koh21a.html.

David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Remi Le Priol, and Aaron Courville. Out-of-distribution generalization via risk extrapolation (REx). In *International Conference on Machine Learning*, pp. 5815–5826, 2021. URL https://proceedings.mlr.press/v139/krueger21a.html.

Bo Li, Yifei Shen, Yezhen Wang, Wenzhen Zhu, Colorado J. Reed, Jun Zhang, Dongsheng Li, Kurt Keutzer, and Han Zhao. Invariant information bottleneck for domain generalization. In *AAAI Conference on Artificial Intelligence*, 2022. URL https://arxiv.org/abs/2106.06333.

Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M. Hospedales. Learning to generalize: Meta-learning for domain generalization. In *AAAI Conference on Artificial Intelligence*, 2018. doi: 10.1609/aaai.v32i1.11596. URL https://ojs.aaai.org/index.php/AAAI/article/view/11596.

Yong Lin, Hanze Dong, Hao Wang, and Tong Zhang. Bayesian invariant risk minimization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16021–16030, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Lin_Bayesian_Invariant_Risk_Minimization_CVPR_2022_paper.html.

Jiashuo Liu, Zheyan Shen, Yue He, Xingxuan Zhang, Renzhe Xu, Han Yu, and Peng Cui. Towards out-of-distribution generalization: A survey, 2023. URL https://arxiv.org/abs/2108.13624. Revised preprint.

Yi Ma, Harm Derksen, Wei Hong, and John Wright. Segmentation of multivariate mixed data via lossy data coding and compression. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(9):1546–1562, 2007. doi: 10.1109/TPAMI.2007.1085. URL https://people.eecs.berkeley.edu/~yima/psfile/Ma-PAMI07.pdf.

Vaishnavh Nagarajan, Anders Andreassen, and Behnam Neyshabur. Understanding the failure modes of out-of-distribution generalization. In *International Conference on Learning Representations*, 2021. URL https://arxiv.org/abs/2010.15775.

Hyeonseob Nam, HyunJae Lee, Jongchan Park, Wonjun Yoon, and Donggeun Yoo. Reducing domain gap by reducing style bias. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8690–8699, 2021. URL https://openaccess.thecvf.com/content/CVPR2021/html/Nam_Reducing_Domain_Gap_by_Reducing_Style_Bias_CVPR_2021_paper.html.

Giambattista Parascandolo, Alexander Neitz, Antonio Orvieto, Luigi Gresele, and Bernhard Schölkopf. Learning explanations that are hard to vary. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=hb1sDDSLbV.

Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. In *International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 39643–39666, 2024. URL https://proceedings.mlr.press/v235/park24c.html.

Kiho Park, Yo Joong Choe, Yibo Jiang, and Victor Veitch. The geometry of categorical and hierarchical concepts in large language models. In *International Conference on Learning Representations*, pp. 76441–76463, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/be7430d22a4dae8516894e32f2fcc6db-Abstract-Conference.html.

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

Alexandre Rame, Corentin Dancette, and Matthieu Cord. Fishr: Invariant gradient variances for out-of-distribution generalization. In *International Conference on Machine Learning*, pp. 18347–18377, 2022. URL https://proceedings.mlr.press/v162/rame22a.html.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do ImageNet classifiers generalize to ImageNet? In *International Conference on Machine Learning*, pp. 5389–5400, 2019. URL https://proceedings.mlr.press/v97/recht19a.html.

Jorma Rissanen. Modeling by shortest data description. *Automatica*, 14(5):465–471, 1978. doi: 10.1016/0005-1098(78)90005-5. URL https://www.sciencedirect.com/science/article/pii/0005109878900055.

Mateo Rojas-Carulla, Bernhard Schölkopf, Richard Turner, and Jonas Peters. Invariant models for causal transfer learning. *Journal of Machine Learning Research*, 19(36):1–34, 2018. URL https://jmlr.org/papers/v19/16-432.html.

Elan Rosenfeld, Pradeep Ravikumar, and Andrej Risteski. The risks of invariant risk minimization. In *International Conference on Learning Representations*, 2021. URL https://arxiv.org/abs/2010.05761.

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://arxiv.org/abs/1911.08731.

Yuge Shi, Jeffrey Seely, Philip H. S. Torr, N. Siddharth, Awni Hannun, Nicolas Usunier, and Gabriel Synnaeve. Gradient matching for domain generalization. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=vDwBW49HmO.

Ha Min Son, Zhe Zhao, Shahbaz Rezaei, and Xin Liu. Domain generalization in-the-wild: Disentangling classification from domain-aware representations. *arXiv preprint arXiv:2508.21769*, 2025. URL https://arxiv.org/abs/2508.21769.

Naftali Tishby, Fernando C. Pereira, and William Bialek. The information bottleneck method. *arXiv preprint arXiv:physics/0004057*, 2000. URL https://arxiv.org/abs/physics/0004057.

Jindong Wang, Cuiling Lan, Chang Liu, Yidong Ouyang, Tao Qin, Wang Lu, Yiqiang Chen, Wenjun Zeng, and Philip S. Yu. Generalizing to unseen domains: A survey on domain generalization. *IEEE Transactions on Knowledge and Data Engineering*, 35(8), 2023. doi: 10.1109/TKDE.2022.3178128. URL https://arxiv.org/abs/2103.03097.

Kai Xiao, Logan Engstrom, Andrew Ilyas, and Aleksander Madry. Noise or signal: The role of image backgrounds in object recognition. In *International Conference on Learning Representations*, 2021. URL https://arxiv.org/abs/2006.09994.

Yaodong Yu, Kwan Ho Ryan Chan, Chong You, Chaobing Song, and Yi Ma. Learning diverse and discriminative representations via the principle of maximal coding rate reduction. *Advances in Neural Information Processing Systems*, 33, 2020.

Yaodong Yu, Sam Buchanan, Druv Pai, Tianzhe Chu, Ziyang Wu, Shengbang Tong, Hao Bai, Yuexiang Zhai, Benjamin D Haeffele, and Yi Ma. White-box transformers via sparse rate reduction: Compression is all there is? *Journal of Machine Learning Research*, 25(300):1–128, 2024.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations*, 2017. URL https://arxiv.org/abs/1611.03530.

Dinghuai Zhang, Kartik Ahuja, Yilun Xu, Yisen Wang, and Aaron Courville. Can subnetwork structure be the key to out-of-distribution generalization? In *International Conference on Machine Learning*, pp. 12356–12367, 2021. URL https://proceedings.mlr.press/v139/zhang21a.html.

Hanlin Zhang, Yi-Fan Zhang, Weiyang Liu, Adrian Weller, Bernhard Schölkopf, and Eric P. Xing. Towards principled disentanglement for domain generalization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. URL https://arxiv.org/abs/2111.13839.

Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*, 2018. URL https://arxiv.org/abs/1710.09412.

Kaiyang Zhou, Yongxin Yang, Yu Qiao, and Tao Xiang. Domain generalization with MixStyle. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=6xHJ37MVxxp.

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. doi: 10.1109/TPAMI.2022.3195549. URL https://arxiv.org/abs/2103.02503.

Xiao Zhou, Yong Lin, Weizhong Zhang, and Tong Zhang. Sparse invariant risk minimization. In *International Conference on Machine Learning*, pp. 27222–27244, 2022. URL https://proceedings.mlr.press/v162/zhou22e.html.

Wei Zhu, Le Lu, Jing Xiao, Mei Han, Jiebo Luo, and Adam P. Harrison. Localized adversarial domain generalization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. URL https://arxiv.org/abs/2205.04114.

A Details of the Motivating Experiment

This appendix provides the complete experimental and computation details for the binary MNIST example in Sections 3.2 and 4.1.

A.1 Data construction

We retain digits 0 and 1 from the first 50,000 MNIST images, giving 10,610 training images divided into two equally sized environments. Each input has two color channels at 14×14 resolution. The digit is rendered as black ink over a red or green background using the original grayscale intensity as opacity, so antialiased digit edges are preserved. Background color is sampled independently with probability 0.5 and is independent of the digit and environment.

The correct-label task uses the original digit identity. In the noisy-label task, each observed training label is independently flipped with probability 0.25. The tasks use exactly the same images, background colors, environment assignments, initialization, minibatch order, and number of optimizer updates. Only the observed training labels differ. All representation, background-exchange, and saliency analyses use training images.

The held-out test set contains 2,055 digit-0 and digit-1 images with their original labels. Every test background is gray: a red or green background pixel $[b, 0]$ or $[0, b]$ is replaced by $[b/2, b/2]$. The two channels are therefore identical, their per-pixel sum is preserved, and the black digit remains unchanged. All models are evaluated on the same gray images. The test set is used only for final evaluation and does not affect training, hyperparameter choice, or checkpoint selection.

A.2 Architecture and optimization

All runs use the same lightweight CNN: three 3×3 convolutions with 16, 32, and 64 channels, ReLU activations, and max pooling after the first two convolutions. A 128-unit ReLU layer feeds a scalar classifier, giving 97,425 parameters. The two control models use IRMv1 as defined in Section 3.1, with penalty weight 1 from the first optimizer update and throughout training; the reference loss-rescaling divisor is therefore also 1. The *LoCaL* run uses the same IRMv1 objective on the noisy-label task with the additional coding-rate penalty, with $\epsilon^2 = 10$ and $\lambda = 100$.

SGD uses momentum 0.9 and global gradient-norm clipping at 1. Minibatches initially contain 128 images from each environment. The learning rate starts at 0.01 and is reduced by a factor of 0.1 after 200, 500, and 800 epochs. After 1,000 epochs, SGD continues on the full training environments at learning rate 0.005, halved every 5,000 updates. The noisy-label IRMv1 model first exceeds 99.9% whole-training accuracy after 6,800 full-environment updates. All three runs use seed 42, the same training phases and minibatch order, and 48,800 total optimizer updates. Checkpoint selection uses training accuracy only.

On the gray-background test set, the correct-label IRMv1 model classifies 2,053 of 2,055 images correctly (99.903%), while the noisy-label IRMv1 model classifies 1,372 correctly (66.764%). The latter accuracy is 40.565% on digit 0 and 91.165% on digit 1, revealing a substantial prediction bias after the colored background cue is removed. With the fixed values $\epsilon^2 = 10$ and $\lambda = 100$, *LoCaL* obtains 90.264% training accuracy and 78.443% gray-background test accuracy.

A.3 Representation spectra

We analyze the 128-dimensional activations after the fully connected ReLU and before classification. Features are raw and uncentered. We define the 75% energy rank as

$$r_{75} = \min \left\{ r : \frac{\sum_{i=1}^{r} \sigma_i^2}{\sum_i \sigma_i^2} \geq 0.75 \right\}. \tag{10}$$

Within each individual environment and after pooling the environments, the energy ranks for digit 0, digit 1, and both digits are $(1, 1, 2)$ with correct-label IRMv1 and $(2, 2, 3)$ with noisy-label IRMv1. For *LoCaL*, all six corresponding class-specific ranks are 1. The 75% threshold is used only as an interpretable summary for comparison; these values should not be read as the exact algebraic ranks

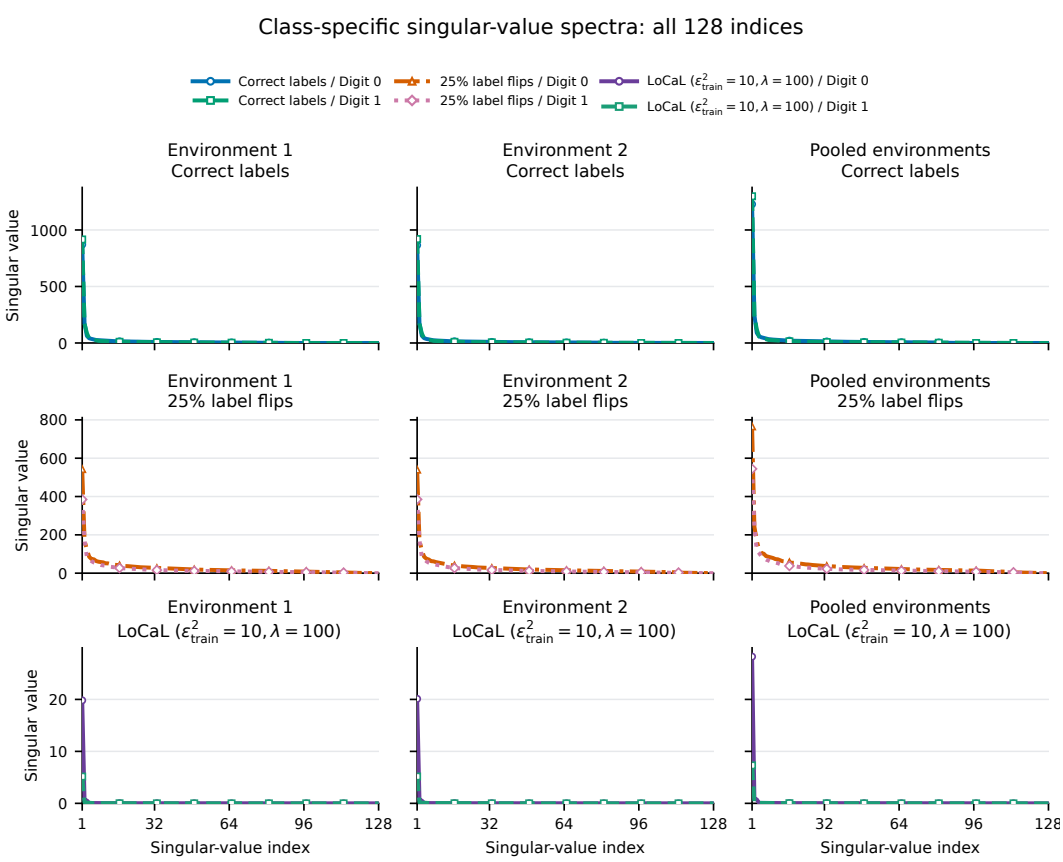

Figure 3: Complete class-specific singular-value spectra. Columns show Environment 1, Environment 2, and the pooled environments; rows show correct-label IRMv1, 25%-flip IRMv1, and *LoCaL* with the fixed values $(\epsilon^2, \lambda) = (10, 100)$. Each panel compares digits 0 and 1 using distinct colors, line styles, and markers. Singular values are computed from raw, uncentered representations, and the linear vertical scale is shared within each row.

or as a complete characterization of representation complexity. Figures 3 and 4 therefore report the complete 128-index spectra on linear and logarithmic axes. Their first two rows show the matched IRMv1 controls, and the third shows *LoCaL* with the fixed values $(\epsilon^2, \lambda) = (10, 100)$.

A.4 Additional diagnostic protocols

For the lossy-coding-rate comparison reported in Section 4.1, we apply Equation (3) with $\epsilon^2 = 10$ to the same raw, uncentered training representations, separately for each environment and digit subset. For the background-exchange diagnostic, we exchange the red and green backgrounds of every training image while preserving its black digit and label, and then measure the fraction of model predictions that change.

For the four training images in Figure 1, SmoothGrad averages signed binary-logit gradients over 64 shared Gaussian perturbations with standard deviation 0.05, after which we take the channelwise absolute maximum. Each nonnegative map is normalized to sum to 100%, and all models share the same perturbations, color mapping, and limits. Cyan contours mark the digit strokes. These maps are qualitative local-sensitivity illustrations; the background-exchange experiment above provides the controlled behavioral diagnostic over the full training set.

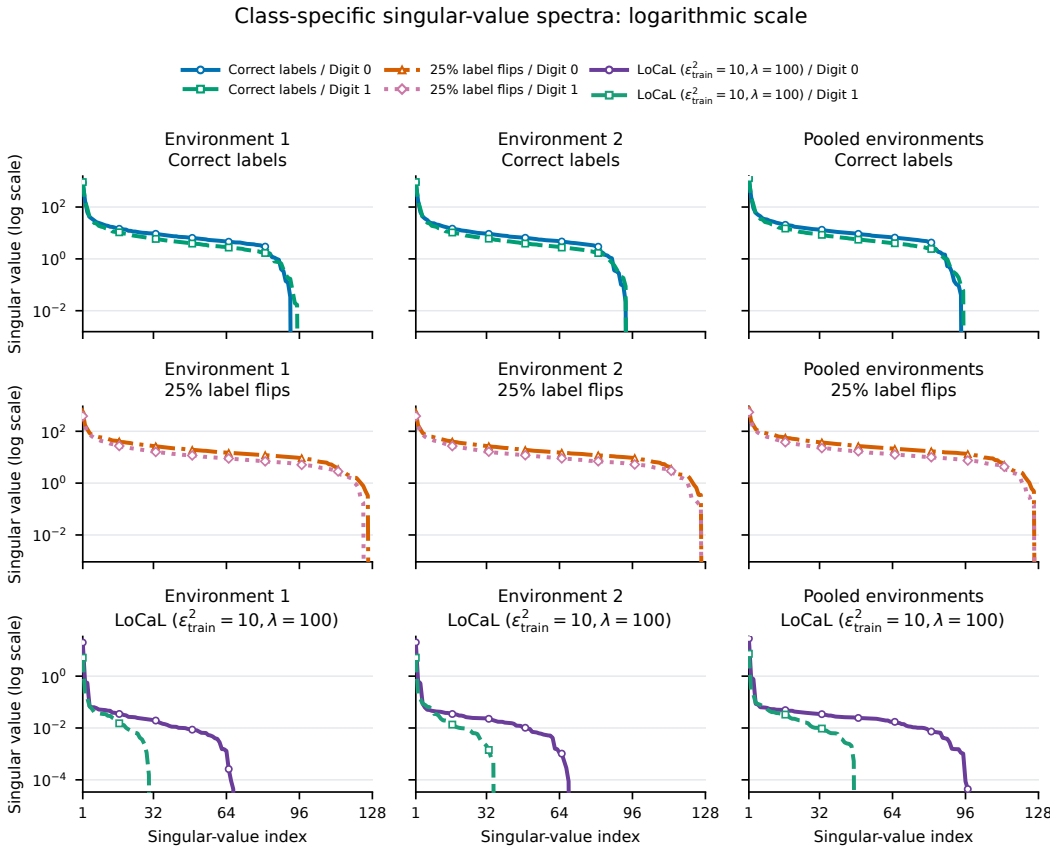

Figure 4: Logarithmic view of the same complete class-specific singular-value spectra as Figure 3. Columns and rows use the same models and subsets. Each row has a shared logarithmic vertical scale spanning six orders of magnitude below its row maximum; the underlying raw, uncentered singular values are unchanged.

B EXPERIMENTAL PROTOCOL AND RESULT PROVENANCE

B.1 REPORTING CONVENTION

The benchmark tables in Sections 4 reproduce the TPAMI means and uncertainty values verbatim. The source manuscript states that performance is averaged over five runs but does not consistently identify whether the displayed uncertainty is a population standard deviation, sample standard deviation, or standard error. We therefore refer to it as *reported uncertainty* rather than assigning a stronger statistical interpretation. No difference is described as statistically significant without an appropriate paired analysis.

Hyperparameter selection uses a 10% split of the training data in the protocols below. The Oracle row is the name used by the source experiments for data without the non-causal correlation; it is not a mathematical upper bound and is not available in a deployment setting.

We set the coding-precision parameter to $\epsilon^2 = 10$ throughout. In the motivating experiment of Sections 3.2 and 4.1, we directly set $\lambda = 100$. In the benchmark experiments below, the coefficient multiplying the coding-rate penalty is selected from the stated training-derived validation grids.

B.2 FULL-AC-CMNIST

Full-AC-CMNIST uses ten digit classes, 20% label noise, and color–label correlations $(0.999, 0.7, 0.1)$ in the two training environments and the test environment. Full-AC-CMNIST-1500 and Full-AC-CMNIST-1000 contain $1{,}500$ and $1{,}000$ training examples, respectively. The model is a three-layer MLP whose two hidden layers have the same width, varied from 64 to 1024.

Oracle, ERM, and plain IRMv1 use Adam with learning rate 10^{-3} and ℓ_2 regularization weight 10^{-4}. IRMv1 uses penalty coefficient 1 for the first 100 steps and 10^4 thereafter, with 501 steps in total. Bayesian IRM trains for $1{,}500$ steps, reduces the learning rate by a factor of 0.1 halfway through training, and changes its invariance coefficient from 0 to 10^4 after 200 steps.

For IB–IRMv1, a 10% training-derived validation split selects the variance coefficient from powers of ten between 10^{-1} and 10^4. MRM first trains the full model, then learns differentiable binary masks with fixed network weights, and finally rewinds and retrains the selected subnetwork. SparseIRM assigns mask scores to all linear-layer weights and biases and selects a retained ratio from $\{0.4, 0.6, 0.8\}$. For IRMv1–*LoCaL*, the same validation fraction selects the coding coefficient from powers of ten between 10^{-2} and 10^3, with invariance coefficient 10^4 and fixed $\epsilon^2 = 10$. The REx variants use a risk-variance coefficient of 10^3. After hyperparameter selection, test performance is evaluated over five runs.

B.3 CIFAR-MNIST AND COLORED OBJECT

CIFAR-MNIST concatenates CIFAR-10 and MNIST components, uses 10% label noise, and has correlation tuple $(0.999, 0.8, 0.1)$. It uses ResNet-18, SGD with momentum 0.9, learning rate 0.01, optimizer weight decay 10^{-3}, batch size 1024, and 500 steps.

Colored Object superimposes eight MSCOCO object classes on colored backgrounds. It uses the same label-noise level and correlation tuple, together with ResNet-18, SGD with momentum 0.9, learning rate 0.01, optimizer weight decay 10^{-3}, batch size 1000, and 800 steps.

For both datasets, the IRMv1 and REx coefficients are 10^4, SparseIRM selects from retained ratios $\{0.4, 0.6, 0.8\}$, and IB and *LoCaL* select their coefficients from $\{10^{-2}, 10^{-1}, 10^0, 10^1, 10^2\}$. For *LoCaL*, ϵ^2 remains fixed at 10.

B.4 CS-MNIST

CS-MNIST has ten digit classes, two training environments of $20{,}000$ examples with color-label correlations 1.0 and 0.9, and a test environment of $20{,}000$ examples whose colors are random. The network has three 3×3 convolutional layers with 64, 128, and 256 channels. Each layer is followed by ReLU and batch normalization; the second convolution uses stride 2. Adaptive average pooling

produces a 256-dimensional representation for a linear classifier. Convolutional and linear weights use Xavier uniform initialization and zero biases.

Optimization uses SGD with initial learning rate 0.1, a decay every 600 steps, batch size 128, 2,000 steps, cross-entropy loss, and no optimizer weight decay. IRMv1, REx, IB, and *LoCaL* select their coefficients from $\{10^{-2}, 10^{-1}, 10^{0}, 10^{1}, 10^{2}\}$; SparseIRM selects from $\{0.4, 0.6, 0.8\}$. For *LoCaL*, ϵ^2 remains fixed at 10.

B.5 Feature-difference visualization

Figure 2 uses the width-1024 Full-AC-CMNIST-1500 models. It changes the color of one digit-0 test image, computes the absolute coordinate-wise difference of the corresponding feature vectors, and reshapes the 1024 values into a 32×32 display. The reshape is a visualization convention and does not impose spatial adjacency on representation coordinates. Because the diagnostic contains one selected image, it is not treated as a population-level measure of non-causal-feature removal.

C Coding-rate derivation and implementation details

C.1 Pooled upper bound for class-wise coding

Let $\boldsymbol{M}_k = \boldsymbol{Z}_k \boldsymbol{Z}_k^\top / m_k$, $p_k = m_k/m$, and $\rho = d/\epsilon^2$. Because $\boldsymbol{Z}$ concatenates the class-specific matrices,

$$\frac{1}{m}\boldsymbol{Z}\boldsymbol{Z}^\top = \sum_k p_k \boldsymbol{M}_k. \tag{11}$$

The map $\boldsymbol{M} \mapsto \frac{1}{2}\log\det(\boldsymbol{I}_d + \rho \boldsymbol{M})$ is concave on the positive-semidefinite cone. Jensen's inequality therefore gives

$$\sum_k p_k \mathcal{R}_\epsilon(\boldsymbol{Z}_k) = \frac{1}{2}\sum_k p_k \log\det(\boldsymbol{I}_d + \rho \boldsymbol{M}_k) \le \frac{1}{2}\log\det\left(\boldsymbol{I}_d + \rho \sum_k p_k \boldsymbol{M}_k\right) = \mathcal{R}_\epsilon(\boldsymbol{Z}). \tag{12}$$

This proves Equation (6). Since the log-determinant is strictly concave on positive-definite matrices, equality holds when all nonempty classes have the same second-moment matrix; otherwise the bound is strict.

C.2 Scale conventions and numerical evaluation

The coding rate changes under feature rescaling, so the feature and classifier scale are part of the training convention. A classifier norm bound $\|\boldsymbol{W}\|_F \le B$ restricts compensating amplification, while head regularization $\eta \|\boldsymbol{W}\|_F^2 / 2$ penalizes it but defines a different optimization problem. Unit-normalized features instead fix total energy and make the coding penalty favor spectral concentration; a trainable score scale or temperature must then also be accounted for. The motivating experiment uses no explicit ℓ_2 regularization, optimizer weight decay, batch normalization, or dropout. The benchmark experiments follow the regularization and optimization conventions reported in Appendix B. We therefore interpret the coding rate as a representation-complexity penalty within each stated training convention, not as a scale-invariant quantity.

For $\alpha = d/(m\epsilon^2)$ and $\boldsymbol{A} = \boldsymbol{I}_d + \alpha \boldsymbol{Z}\boldsymbol{Z}^\top = \boldsymbol{L}\boldsymbol{L}^\top$, Cholesky evaluation gives $\mathcal{R}_\epsilon(\boldsymbol{Z}) = \sum_j \log L_{jj}$. The equivalent sample-space form, $\frac{1}{2}\log\det(\boldsymbol{I}_m + \alpha \boldsymbol{Z}^\top \boldsymbol{Z})$, uses the same α. Centering, normalization, and weighting remain in the gradient computation; IRMv1 also requires differentiating through its squared scalar derivative. Numerical precision, any added jitter, penalty schedules, and checkpoint selection should be documented alongside the feature location and classifier scale.

C.3 Numerical accuracy of the variance approximation

We illustrate the scale of the first-order regime numerically in Figure 5. We generate 100 centered Gaussian representation matrices $\boldsymbol{Z} \in \mathbb{R}^{128\times 1024}$ and rescale each so that $\mathrm{tr}(\boldsymbol{M}_{\boldsymbol{Z}}) = 128$. At $\epsilon^2 = 10$, the first-order variance term overestimates the exact coding rate by a median of 398%. Reducing the median relative error below 10%, 5%, and 1% requires ϵ^2 to reach approximately

690, 1,409, and 7,172, respectively. Thus, even for simple random representations, the variance approximation becomes accurate only when ϵ^2 is very large relative to the feature spectrum.

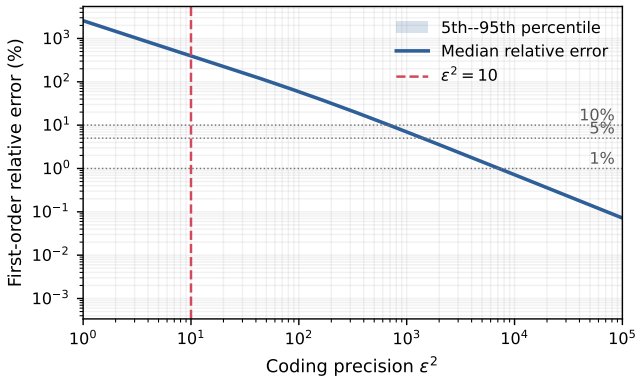

Figure 5: Relative error of the first-order variance approximation on random representations. Each of the 100 trials uses a centered Gaussian matrix $\boldsymbol{Z} \in \mathbb{R}^{128\times1024}$ normalized so that $\mathrm{tr}(\boldsymbol{M}_{\boldsymbol{Z}}) = 128$. The curve and shaded region show the median and 5th–95th percentiles, respectively.