Stabilizing Variational Information Bottleneck Under Covariance Shift

Anonymous authors
Paper under double-blind review

Abstract

In the Gaussian Information Bottleneck (IB) setting, the learned representation depends on input covariance statistics that can be distorted by finite-sample estimation error and distribution shift. We show that such perturbations can destabilize the Gaussian IB eigenstructure. Motivated by this sensitivity, we study how covariance shift distorts, and we propose two regularizers to improve robustness. *Shrinkage-VIB* encourages isotropic latent covariance via a Ledoit–Wolf-inspired penalty, while *Sensitivity-VIB* penalizes encoder Jacobian norm through a finite-difference proxy with a theory-motivated weight scaling. Experiments on MNIST, CIFAR-10, and STL-10 under two covariance-shift protocols show that Sensitivity-VIB is stronger under compound and severe shift, whereas Shrinkage-VIB provides consistent gains under milder geometric shift. Both methods substantially outperform Standard VIB and unregularized baselines.

1 Introduction

The Information Bottleneck principle (1) learns representations Z that preserve information about a target Y while compressing the input X. In the Gaussian setting (23), the optimal linear encoder is determined by the eigenvectors of $M = \Sigma_{X|Y}\Sigma_X^{-1}$, and practical deep instantiations such as Variational IB (VIB) (3) optimize an ELBO-style objective via reparameterization.

A hidden source of brittleness is that these methods depend on input covariance statistics that are estimated from finite data and can also change under distribution shift. In the Gaussian case, perturbations of Σ_X propagate into the eigen-structure of M, making some directions much more sensitive than others. This sensitivity is especially pronounced when eigenvalues approach critical thresholds, as we quantify in Section 3. Such fragility is practically relevant in medical sensing, wearable devices, and industrial monitoring, where feature correlations vary across subjects or sessions while the input–output mechanism remains stable.

We analyze the VIB objective under bounded Frobenius covariance perturbations and derive an upper bound on its worst-case variation (Proposition 1). This bound motivates penalizing the encoder Jacobian norm as a principled surrogate for reducing covariance-shift sensitivity, implemented via a finite-difference proxy justified by Theorem 3, with a theory-motivated weight scaling intended to avoid representation collapse.

Contributions.

- We analyze the Gaussian IB eigen-structure under covariance perturbations and quantify how critical compression thresholds become unstable as eigenvalues approach unity (Section 3).
- We derive an upper bound on worst-case VIB loss variation under bounded Frobenius covariance perturbations, and identify the encoder Jacobian norm as a principled surrogate minimizer (Proposition 1, Theorem 3).
- We propose **Shrinkage-VIB**, a lightweight latent covariance regularizer, and **Sensitivity-VIB**, a finite-difference Jacobian-norm penalty with theory-motivated weight scaling intended to increase robustness (Section 4).
- We evaluate on MNIST, CIFAR-10, and STL-10 under two covariance-shift protocols, showing that Sensitivity-VIB is the dominant method under compound and severe shift while Shrinkage-VIB remains competitive under milder geometric shift (Section 5).

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.

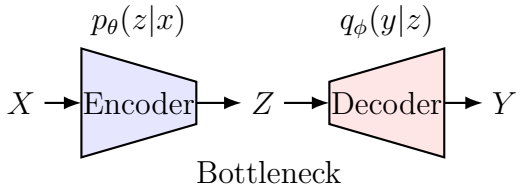


Figure 1: Information Bottleneck diagram: compression from X to Z, followed by relevance mapping from Z to Y.

2 Related Work

Information bottleneck. Tishby et al. (1) introduced IB as a rate-distortion problem over the relevance variable Y. Chechik et al. (23) derived the closed-form Gaussian solution, whose eigen-structure we analyze under covariance perturbations. Alemi et al. (3) proposed VIB, a variational deep-learning instantiation that we extend with covariance-robust regularizers.

Distributionally robust optimization (DRO). DRO (4) minimizes worst-case loss over an ambiguity set. Our first-order bound (Proposition 1) admits a DRO interpretation specialized to a Frobenius covariance uncertainty set within the IB objective. Unlike iterative PGD approaches (5), our first-order analysis yields a closed-form worst-case perturbation direction and a single tractable regularizer, requiring no inner-loop optimization.

Domain generalization. IRM (6) encourages representations whose optimal classifier is invariant across environments. CORAL (7) and related covariance-alignment methods minimize between-domain covariance mismatch, but require multiple labeled environments at training time. Sensitivity-VIB instead targets robustness to covariance perturbations within a single training distribution, requiring no target-domain labels or environment annotations.

Jacobian and smoothness regularization. Sokolić et al. (21), Hoffman et al. (9) and Varga et al. (10) propose Jacobian-norm penalties to improve adversarial robustness in discriminative classifiers. Rifai et al. (11) penalize the encoder Jacobian in contractive autoencoders to encourage locally invariant representations. Sensitivity-VIB derives a related penalty from a first-order bound on worst-case VIB loss variation under covariance shift (Proposition 1, Theorem 3), giving a distinct information-theoretic motivation and a theory-prescribed weight scaling, rather than a margin or reconstruction objective.

Covariance estimation. Ledoit and Wolf (22) derive an analytically optimal shrinkage estimator for the sample covariance matrix. Shrinkage-VIB is complementary in spirit: rather than improving the covariance estimate used as input to the model, it regularizes the geometry of the learned latent representation directly, encouraging isotropy as a defense against covariance-shift propagation into the bottleneck.

3 Theoretical Background: IB Under Covariance Perturbations

3.1 Information Bottleneck: Definition and System Model

Let (X, Y) be jointly distributed random variables with joint law $p(x, y)$. The Information Bottleneck (IB) framework aims to construct a compressed representation Z of X that preserves as much information as possible about Y. The encoder is modeled by a conditional distribution $p_\theta(z|x)$, and the Markov chain $Y - X - Z$ is assumed to hold. The classical IB optimization problem is

$$\max_{p_\theta(z|x)} \; I(Z;Y) - \beta I(X;Z), \tag{1}$$

where $\beta \geq 0$ is a tradeoff parameter that controls the balance between relevance and compression. Equivalently, one may write

$$\min_{p_\theta(z|x)} \; I(X;Z) - \beta I(Z;Y).$$

A larger value of β places greater emphasis on preserving information about Y, whereas a smaller value promotes stronger compression of X. In general, the encoder $p_\theta(z|x)$ is chosen so that Z remains highly informative about Y while discarding irrelevant details of X. When (X, Y) are jointly Gaussian and the encoder is restricted to be Gaussian, the IB problem admits a closed-form solution.

3.2 Gaussian IB and Its Eigen-Structure

Let (X, Y) be jointly Gaussian with $X \in \mathbb{R}^d$. In the Gaussian IB setting, the optimal encoder is a noisy linear projection of the form

$$Z = AX + \xi, \qquad \xi \sim \mathcal{N}(0, I), \tag{2}$$

where A is the projection matrix and ξ is independent Gaussian noise. The relevant eigen-structure is determined by

$$M \triangleq \Sigma_{X|Y}\Sigma_X^{-1}, \tag{3}$$

where Σ_X is the covariance matrix of X and $\Sigma_{X|Y}$ is the conditional covariance of X given Y.

Theorem 1 (Optimal Gaussian IB projection (23)). *For the Gaussian Information Bottleneck problem, the optimal encoder in equation 2 has a projection matrix of the form*

$$A = \begin{cases} [\,0^\top, \ldots, 0^\top\,], & 0 \le \beta \le \beta_1^c, \\ [\,\alpha_1 v_1^\top,\, 0^\top,\, \ldots,\, 0^\top\,], & \beta_1^c \le \beta \le \beta_2^c, \\ [\,\alpha_1 v_1^\top,\, \alpha_2 v_2^\top,\, 0^\top,\, \ldots,\, 0^\top\,], & \beta_2^c \le \beta \le \beta_3^c, \\ \vdots \end{cases} \tag{4}$$

where $\{v_i\}$ are the eigenvectors of $\Sigma_{X|Y}\Sigma_X^{-1}$ ordered by ascending eigenvalue, $0 \le \lambda_1 \le \lambda_2 \le \cdots \le \lambda_d$, $\beta_i^c \triangleq 1/(1-\lambda_i)$, and $\alpha_i = \sqrt{\frac{\beta(1-\lambda_i)-1}{\lambda_i r_i}}$, $r_i = v_i^\top \Sigma_X v_i$.

Thus, as β increases, the optimal encoder activates the eigen-directions one by one in ascending order of the eigenvalues λ_i (23). For the linear-Gaussian model, the IB objective (GIB) can also be written in log-determinant form as $\mathcal{L}_{\text{GIB}} = (1-\beta)\log\left|A\Sigma_X A^\top + \Sigma_\xi\right| - \log|\Sigma_\xi| + \beta\log\left|A\Sigma_{X|Y}A^\top + \Sigma_\xi\right|$. Up to additive constants, this expression captures the tradeoff between preserving information about Y and compressing X through the Gaussian encoder in equation 2.

Remark 1 (Gaussian IB specialization). *In the Gaussian IB setting, the relevant matrix is $M = \Sigma_{X|Y}\,\Sigma_X^{-1}$. Therefore, under the covariance-shift model in Definition 1, the perturbation acts through $\Sigma_X^{(\rho)} = \Sigma_X^{(0)} + \Delta_\rho$, while $\Sigma_{Y|X}$ is held fixed. This is the precise sense in which the covariance perturbation changes the IB eigen-structure. Now, we can write $M^{(\rho)} := \Sigma_{X|Y}\big(\Sigma_X^{(\rho)}\big)^{-1}$, $M^{(0)} := \Sigma_{X|Y}\big(\Sigma_X^{(0)}\big)^{-1}$. Assuming $\Sigma_X^{(0)}$ is invertible and $\|\Delta_\rho\|$ is sufficiently small, a first-order expansion gives*

$$M^{(\rho)} = \Sigma_{X|Y}\big(\Sigma_X^{(0)} + \Delta_\rho\big)^{-1} = M^{(0)} - \Sigma_{X|Y}\big(\Sigma_X^{(0)}\big)^{-1}\Delta_\rho\big(\Sigma_X^{(0)}\big)^{-1} + O(\|\Delta_\rho\|^2). \tag{5}$$

Hence the induced perturbation $\Delta M := M^{(\rho)} - M^{(0)}$ is controlled by Δ_ρ.

In practice one works with $\Sigma_X^{(\rho)} = \Sigma_X^{(0)} + \Delta_\rho$, where Δ_X arises from finite-sample estimation, model mismatch, or distribution shift. The resulting perturbation of M propagates through three mechanisms.

3.3 Covariance perturbation model

Definition 1 (Covariance-shift perturbation). *Let P_0 denote the source distribution of (X, Y), with $\mu_X^{(0)} := \mathbb{E}_{P_0}[X]$, $\Sigma_X^{(0)} := \mathrm{Cov}_{P_0}(X) = \mathbb{E}_{P_0}\big[(X - \mu_X^{(0)})(X - \mu_X^{(0)})^\top\big]$. For perturbation factor of $\rho \in [0, 1]$, we say that P_ρ is a* covariance perturbation *of P_0 if the conditional mechanism is preserved while the input covariance changes, namely*

$$P_\rho(Y \mid X) = P_0(Y \mid X), \quad \Sigma_X^{(\rho)} = \Sigma_X^{(0)} + \Delta_\rho, \quad \Delta_\rho = \Delta_\rho^\top, \quad \|\Delta_\rho\|_F \le \epsilon(\rho), \tag{6}$$

where $||.||_F$ is Frobenius norm. Equivalently, the shift acts only on the second-order statistics of X, while the input–output relation remains unchanged.

Definition 2 (Linear shift model used in experiments)**.** *A convenient concrete model is* $X_\rho = A_\rho X + \xi_\rho$, $\mathbb{E}[\xi_\rho] = 0$, $\mathrm{Cov}(\xi_\rho) = \Sigma_\xi(\rho)$, *which induces*

$$\Sigma_X^{(\rho)} = A_\rho \Sigma_X^{(0)} A_\rho^\top + \Sigma_\xi(\rho).$$

When $P_\rho(Y \mid X) = P_0(Y \mid X)$*, this is a covariance shift in the sense of Definition 1.*

The eigen-structure described above can become unstable under small perturbations of the input covariance. Intuitively, when an eigenvalue λ_i is close to its critical value of one, the corresponding compression threshold β_i^c becomes highly sensitive to even minor changes in the covariance statistics. Similarly, when two consecutive eigenvalues are nearly equal (i.e., a small spectral gap), the corresponding encoder directions can rotate substantially under a tiny perturbation. Both effects are classical consequences of Weyl's inequality and the Davis–Kahan theorem (13; 14). Both effects imply that finite-sample estimation error or mild distribution shift can significantly alter which directions the encoder retains and how it compresses the input. We study these instabilities formally in Appendix B, and they directly motivate the robustness analysis and regularizers developed in the following sections.

4 Proposed Robust Loss Functions

4.1 Standard VIB baseline:

The standard VIB loss is

$$\mathcal{L}_{\mathrm{VIB}} = \mathbb{E}\big[-\log q_\phi(y \mid z)\big] + \beta\, \mathbb{E}\big[D_{\mathrm{KL}}(p_\theta(z \mid x) \,\|\, p_\theta(z))\big], \tag{7}$$

Let the encoder $p_\theta(z \mid x) = \mathcal{N}(\mu_\theta(x), \mathrm{diag}(\sigma_\theta^2(x)))$ map each input $x \in \mathbb{R}^d$ to a Gaussian latent variable $z \in \mathbb{R}^m$, with prior $p_\theta(z) = \mathcal{N}(0, I_m)$. Using the reparameterization $z = \mu_\theta(x) + \sigma_\theta(x) \odot \varepsilon$, $\varepsilon \sim \mathcal{N}(0, I_m)$, and writing $s_j(x) := \log \sigma_{\theta,j}^2(x)$ for numerical stability, the closed-form KL divergence between $\mathcal{N}(\mu_\theta(x), \mathrm{diag}(e^{s_\theta(x)}))$ and $\mathcal{N}(0, I_m)$ gives the explicit VIB loss

$$\mathcal{L}_{\mathrm{VIB}}(\theta, \phi; x, y) = \underbrace{\mathbb{E}_\varepsilon\Big[-\log q_\phi\big(y \mid z\big)\Big]}_{\text{reconstruction term}} + \underbrace{\frac{\beta}{2} \sum_{j=1}^{m} \Big(e^{s_{\theta,j}(x)} + \mu_{\theta,j}^2(x) - 1 - s_{\theta,j}(x)\Big)}_{\text{KL term}}. \tag{8}$$

Each KL term $e^{s_j} + \mu_j^2 - 1 - s_j \geq 0$ by the inequality $t - 1 - \log t \geq 0$ for $t > 0$, with equality if and only if $\mu_j = 0$ and $s_j = 0$.

4.2 Sensitivity-VIB: Method and Implementation

Explicit VIB loss in terms of μ and σ. Let $\Sigma_X^{(0)}$ denote the batch input covariance and define the Frobenius uncertainty set

$$\mathcal{U} = \big\{\Delta : \|\Delta\|_F \leq \epsilon,\ \Sigma_X^{(\rho)} = \Sigma_X^{(0)} + \Delta \succ 0\big\}.$$

The robust Covariance-Aware Variational Information Bottleneck (CA-VIB) objective is

$$\min_{\theta,\phi}\ \sup_{\Delta \in \mathcal{U}}\ \mathcal{L}_{\mathrm{VIB}}(\theta, \phi; \Sigma_X^{(\rho)} = \Sigma_X^{(0)} + \Delta_\rho), \tag{9}$$

where θ are encoder parameters and ϕ are classifier parameters.

Pointwise input perturbation and sensitivity analysis. We consider a pointwise perturbation of the input:

$$x \ \longmapsto\ x + \delta, \qquad \delta \in \mathbb{R}^d, \qquad \|\delta\|_2 \leq \epsilon. \tag{10}$$

Let $J_\mu(x) := \nabla_x \mu_\theta(x) \in \mathbb{R}^{m \times d}$ and $J_s(x) := \nabla_x s_\theta(x) \in \mathbb{R}^{m \times d}$ denote the Jacobians of the encoder mean and log-variance with respect to the input. By Taylor's theorem and the chain rule,

$$\mathcal{L}_{\mathrm{VIB}}(\theta, \phi; x + \delta, y) = \mathcal{L}_{\mathrm{VIB}}(\theta, \phi; x, y) + \langle \nabla_x \mathcal{L}_{\mathrm{VIB}}(x, y),\, \delta \rangle + \mathcal{O}(\|\delta\|_2^2),$$

where x is the input and y is the label, the full input gradient decomposes via the chain rule as

$$\nabla_x \mathcal{L}_{\mathrm{VIB}}(x, y) = J_\mu(x)^\top \nabla_\mu \mathcal{L}_{\mathrm{VIB}} + J_s(x)^\top \nabla_s \mathcal{L}_{\mathrm{VIB}}, \tag{11}$$

with $\nabla_\mu \mathcal{L}_{\mathrm{VIB}} \in \mathbb{R}^m$ and $\nabla_s \mathcal{L}_{\mathrm{VIB}} \in \mathbb{R}^m$ given by equation 22–equation 23.

Proposition 1 (First-order perturbation bound)**.** *Under the pointwise perturbation equation 10,*

$$\sup_{\|\delta\|_2 \le \epsilon} \mathcal{L}_{\mathrm{VIB}}(\theta, \phi; x+\delta, y) \;\le\; \mathcal{L}_{\mathrm{VIB}}(\theta, \phi; x, y) \;+\; \epsilon\Big(\big\|\nabla_\mu \mathcal{L}_{\mathrm{VIB}}\big\|_2 \, \|J_\mu(x)\|_{\mathrm{op}} + \big\|\nabla_s \mathcal{L}_{\mathrm{VIB}}\big\|_2 \, \|J_s(x)\|_{\mathrm{op}} \Big) + \mathcal{O}(\epsilon^2). \tag{12}$$

Proof. We will prove it in Appendix equation C. □

Motivated surrogate regularizer. Proposition 1 shows that the worst-case first-order loss increase is controlled by $\epsilon\|\nabla_x \mathcal{L}_{\mathrm{VIB}}\|_2$, which in turn depends on both $\|J_\mu\|_{\mathrm{op}}$ and $\|J_s\|_{\mathrm{op}}$. Computing these operator norms exactly requires full Jacobians, which is expensive. We therefore propose the following *motivated surrogate regularizer*, which penalizes the dominant μ-sensitivity term via a cheap Monte Carlo finite-difference estimate:

$$\widehat{\mathcal{S}}(x;\theta) \;:=\; \frac{\big\|\mu_\theta(x+\epsilon_0\,\varepsilon) - \mu_\theta(x)\big\|_2^2}{\epsilon_0^2}, \;\; \varepsilon \sim \mathcal{N}(0, I_d),$$

with $\epsilon_0 > 0$ a small finite-difference step size. By the limit of $\mathbb{E}_\varepsilon\big[\widehat{\mathcal{S}}(x;\theta)\big] \xrightarrow{\epsilon_0 \to 0} \|J_\mu(x)\|_F^2$, confirming that $\widehat{\mathcal{S}}$ is an estimator of $\|J_\mu\|_F^2$ in the small-step limit. Moreover, $\widehat{\mathcal{S}}$ is a Monte Carlo finite-difference approximation of the local encoder sensitivity. The J_s term in equation 12 is omitted here in terms of practical implementation for tractability; however, it is second-order for small β and contributes negligibly in the regime $\beta \ll 1$ used in our experiments. The Theorem equation 3 in Appendix equation D justifies this choice.

The **Sensitivity-VIB** loss is therefore defined as

$$\mathcal{L}_{\mathrm{SensVIB}}(\theta, \phi; x, y) \;=\; \mathcal{L}_{\mathrm{VIB}}(\theta, \phi; x, y) \;+\; \lambda\, \widehat{\mathcal{S}}(x;\theta), \tag{13}$$

with $\lambda > 0$ a tunable weight proportional to the perturbation radius ϵ. Unlike a fully proved min-max robustness theorem, equation 13 is a principled surrogate: it is motivated by the first-order bound equation 12, targets the dominant source of input sensitivity in the encoder, and is computationally tractable. We note that the J_s contribution, while omitted for tractability, could be incorporated via a second finite-difference call on the log-variance head if stricter robustness guarantees are required.

The Sensitivity-VIB loss equation 13 and Shrinkage VIB loss equation 15 are, to our knowledge, the first VIB loss function derived from a principled min-max covariance perturbation analysis. It differs from prior Jacobian regularizers in that (a) the penalty target is the gradient of the *information-theoretic* VIB objective with respect to the input covariance—not a classifier margin; (b) The regularization weight has a closed-form theoretical prescription.

4.3 Shrinkage VIB: Method and Implementation

Since Sensitivity-VIB shows that the latent representation can be affected by small input perturbations, we are motivated to regularize the covariance of the latent space as a way to mitigate covariance shift in the input data. This leads us to define the Shrinkage-VIB. Shrinkage VIB penalizes the discrepancy between the latent covariance and its isotropic projection:

$$\mathcal{R}_{\mathrm{shrink}}(\mu) = \big\|C(\mu) - \tfrac{\mathrm{tr}C(\mu)}{m}\,\mathbf{I}\big\|_F^2, \tag{14}$$

where $C(\mu)$ is the minibatch latent covariance. This is a *post-hoc latent-space* regularizer, motivated by the Ledoit–Wolf covariance estimator (22): it encourages the learned representation to have isotropic spread, but does not penalize the *encoder's sensitivity* to input-space covariance perturbations. Based on the property of the Shrinkage VIB, we define the loss function as

$$\mathcal{L}_{\mathrm{SensVIB}}(\theta, \phi; x, y) \;=\; \mathcal{L}_{\mathrm{VIB}}(\theta, \phi; x, y) \;+\; \lambda\, \mathcal{R}_{\mathrm{shrink}}(\mu). \tag{15}$$

The mathematical foundations of Shrinkage VIB, including the geometric interpretation of the latent covariance penalty via isotropic projection and eigenvalue dispersion, are presented in Appendix E.

Table 1: Classification accuracy under covariance shift, rotation and translation, CSP1, on MNIST, CIFAR-10, and STL-10.

Method	MNIST				CIFAR-10				STL-10			
	Clean (ρ=0)	Moderate (ρ=0.50)	Strong (ρ=1.00)	Mean	Clean (ρ=0)	Moderate (ρ=0.50)	Strong (ρ=1.00)	Mean	Clean (ρ=0)	Moderate (ρ=0.50)	Strong (ρ=1.00)	Mean
Raw CNN	0.9810	0.9739	0.9387	0.9679	0.7681	0.5682	0.5049	0.5864	0.5259	0.3884	0.3321	0.4065
PCA + LR	0.9063	0.7945	0.4692	0.7346	0.3770	0.2858	0.2506	0.2979	0.3493	0.2951	0.2642	0.3005
Std VIB	0.9883	0.9827	0.9592	0.9787	0.7462	0.5619	0.4629	0.5744	0.5639	0.5032	0.4566	0.5054
Shrinkage VIB	0.9879	0.9861	0.9682	0.9825	0.7856	0.5777	0.4997	0.6006	**0.6519**	**0.5885**	**0.5530**	**0.5916**
Sensitivity VIB	**0.9916**	**0.9876**	**0.9737**	**0.9859**	**0.7876**	**0.6114**	**0.5452**	**0.6229**	0.6375	0.5646	0.5016	0.5669

5 Experiments

5.1 Setup

Datasets. We evaluate on three standard image-classification benchmarks spanning a range of difficulty:

MNIST (15): 28×28 grayscale, 10 classes. We use the standard 60 000 / 10 000 train/test split, reserving 10% of the training set as a validation split (≈54 000 train / 6 000 val).

CIFAR-10 (16): 32×32 RGB, 10 classes. Standard 50 000 / 10 000 split with 10% held out for validation (≈45 000 train / 5 000 val).

STL-10 (17): 96×96 RGB, 10 classes. Following the official split we use the first 5 000 images for training (15% carved out as validation, ≈4 250 train / 750 val) and the remaining 8 000 images as the test set.

Compute resources. All experiments were conducted on the High Performance Computing (HPC) cluster at Oregon State University. Each training run took approximately 15–20 minutes for MNIST, 40–60 minutes for CIFAR-10, and 60–90 minutes for STL-10, with a total estimated compute of approximately 6 GPU-hours across all reported experiments.

General Covariance shift protocol. In natural images, neighboring pixels exhibit strong local dependencies, so the covariance structure of an image is shaped more by geometry than by individual pixel intensities. Motivated by this fact, we model covariance shift through controlled geometric transformations that alter spatial correlations while preserving the underlying semantic content. We evaluate under two test-time covariance-shift protocols (CSP). CSP1 applies random rotations and translations to induce purely geometric covariance change. CSP2 extends CSP1 with additional contrast, noise, and channel scaling perturbations. Full details, including parameter ranges and implementation, plus more details, are given in Appendix F.

Per-dataset accuracy and covariance stability plots are provided in Appendix G for a more detailed view of the results. Sensitivity-VIB and Shrinkage-VIIB (both ours) maintains the highest accuracy at strong shift under CSP2 on all datasets. Moreover, we plot the latent covariance stability, measured by the Frobenius distance, $\Delta_\mu(\rho) = \|\widehat{C}_{\text{train}} - \widehat{C}_{\text{test}}(\rho)\|_F$ vs. shift strength ρ for Standard VIB, Shrinkage VIB, and Sensitivity-VIB. Top row: CSP1. Bottom row: CSP2. Sensitivity-VIB exhibits the smallest latent drift in most settings; absolute drift grows with input resolution (MNIST < CIFAR-10 < STL-10). The Sensitivity-VIB wins MNIST and the CIFAR-10, but the shrinkage wins for the STL-10 dataset. Also, Representative training and test images under both protocols are shown in Appendix H,

5.2 Results Analysis

Tables 1 and 2 and Figures 5, 6, and 7 report classification accuracy under CSP1 and CSP2 across all methods and datasets. Figures 8, 9, and 10 reports latent covariance stability. Broadly, Raw CNN degrades sharply under stronger shift, PCA + logistic regression (PCA (18) followed by logistic regression (19)), we call PCA+LR, is more stable but uniformly weak, Standard VIB and Shrinkage VIB improve robustness modestly, and Sensitivity-VIB achieves the strongest performance at the highest shift levels under the harsher protocol.

Table 2: Classification accuracy under covariance shift, rotation, translation, geometric and photometric perturbations, (CSP2), on MNIST, CIFAR-10, and STL-10.

Method	MNIST			CIFAR-10			STL-10		
	Clean (ρ=0)	Moderate (ρ=0.50)	Strong (ρ=1.00)	Clean (ρ=0)	Moderate (ρ=0.50)	Strong (ρ=1.00)	Clean (ρ=0)	Moderate (ρ=0.50)	Strong (ρ=1.00)
Raw CNN	0.9878	0.7126	0.1451	0.7681	0.2527	0.1273	0.5919	0.2269	0.1196
PCA + LR	0.8972	0.7675	0.4267	0.3779	0.2928	0.2535	0.3493	0.3024	**0.2566**
Std VIB	0.9878	0.9035	0.2154	0.7462	0.2705	0.1590	0.4481	0.2057	0.1214
Shrinkage VIB	**0.9907**	0.9044	0.1460	**0.7876**	0.2176	0.1291	0.5654	0.1650	0.1170
Sensitivity VIB	0.9872	**0.9793**	**0.8668**	0.7856	**0.4162**	**0.2068**	**0.6714**	**0.4520**	0.2355

The covariance-stability plots are consistent with the accuracy trends, with Sensitivity-VIB showing the smallest latent drift in most settings.

Clean accuracy on CIFAR-10 and STL-10 is modest relative to state-of-the-art results on these benchmarks. This is expected under our intentionally constrained experimental setup: a shallow three-layer convolutional stem, latent dimension 16, 50 training epochs, and batch size 128. These design choices are deliberate: our goal is not to maximize clean accuracy, but to study the robustness of the VIB objective under covariance shift in a controlled and reproducible setting. A more expressive backbone would likely improve absolute accuracy, but the comparisons in this paper are made under the same architecture and training budget for all methods, so the results remain directly comparable. Our main conclusions concern relative robustness under the chosen protocol, not absolute benchmark performance.

Raw CNN under shift. The Raw CNN achieves competitive clean accuracy across all datasets but degrades rapidly with ρ. Under the harsher CSP2 protocol (Table 2), strong-shift accuracy falls to 0.1451, 0.1273, and 0.1196 on MNIST, CIFAR-10, and STL-10, respectively, suggesting that an unregularized discriminative model provides little protection against geometric covariance shift regardless of input resolution. Under the milder CSP1 (Table 1), degradation is less severe but still substantial at $\rho = 1.00$ (0.939, 0.5049, and 0.332 on MNIST, CIFAR-10, and STL-10, respectively).

PCA+LR degrades gracefully but is uniformly weak. PCA + logistic regression (PCA (18) followed by logistic regression (19)) PCA+LR retains moderate clean accuracy on MNIST under both protocols and is relatively more stable than Raw CNN at strong shift under CSP2, which may reflect the tendency of principal components to suppress high-frequency geometric variation. However, on CIFAR-10 and STL-10 its clean accuracy is already far below all neural methods, and strong-shift accuracy remains comparable to or worse than Raw CNN , making it an unreliable baseline at higher input resolutions.

Standard VIB provides meaningful compression-based robustness. Standard VIB yields consistent improvement over both baselines in mean accuracy across all datasets and protocols, consistent with the KL bottleneck discarding covariance-sensitive directions in the latent space. The gain is most pronounced on STL-10 under CSP1, where Standard VIB achieves a mean accuracy of 0.505 versus 0.407 for Raw CNN, suggesting that latent compression may be particularly beneficial at higher input resolution.

Shrinkage VIB improves over Standard VIB but plateaus at strong shift. Shrinkage VIB achieves moderate gains over Standard VIB across most settings, consistent with its latent isotropy penalty encouraging more uniform variance allocation across latent dimensions. Notably, on STL-10 under CSP1, Shrinkage VIB achieves the highest mean accuracy among all methods (0.592 vs. 0.567 for Sensitivity-VIB) and also leads at strong shift (0.553 vs. 0.502), indicating that the two methods are competitive and that Shrinkage VIB can be preferable in the mild geometric shift regime. However, Shrinkage VIB's advantage does not extend to the harsher CSP2 protocol, where it underperforms even Standard VIB on MNIST at $\rho = 1.00$ (0.146 vs. 0.215). This is consistent with Remark 4, which shows that penalizing minibatch latent covariance does not bound the encoder Jacobian $\|J_{\mu_\theta}(x)\|_F$ and therefore does not directly constrain input-space sensitivity.

Sensitivity-VIB achieves the strongest performance at high shift under CSP2. Under CSP2 (Table 2), Sensitivity-VIB attains the highest strong-shift accuracy on all three datasets, leading Shrinkage VIB by +72.1 pp on MNIST (0.867 vs. 0.146), +7.8 pp on CIFAR-10, and +11.9 pp on STL-10 at $\rho = 1.00$.

Under the milder CSP1 (Table 1), Sensitivity-VIB remains best on MNIST and CIFAR-10 at a strong shift, but Shrinkage VIB leads on STL-10 in both mean and strong-shift accuracy. Taken together, these results are consistent with the Jacobian-norm penalty suppressing the encoder's input-to-latent sensitivity most effectively when the perturbation is compounded by photometric distortions, while the two methods remain competitive under purely geometric shift on the hardest dataset.

Latent covariance stability corroborates the accuracy results. Figures 8, 9, and 10 reports the Frobenius norm of

$$\Delta_\mu(\rho) = \|\widehat{C}_{\text{train}} - \widehat{C}_{\text{test}}(\rho)\|_F$$

for the three VIB variants under both protocols. Sensitivity-VIB exhibits the smallest covariance drift in most settings, consistent with the accuracy trends. Absolute drift grows with input resolution (MNIST $<$ CIFAR-10 $<$ STL-10). Shrinkage VIB, despite explicitly targeting latent covariance geometry during training, does not achieve comparable stability in most settings, consistent with Remark 4.

Training dynamics. All VIB variants converge smoothly within 50 epochs across all three datasets. Sensitivity-VIB reaches a slightly lower clean validation accuracy than Standard VIB and Shrinkage VIB in early epochs, consistent with the Jacobian penalty trading a small amount of clean-set fit for improved robustness under shift. No evidence of representation collapse is observed at $\lambda_{\text{sens}} = 10^{-3}$, consistent with the weight-scaling analysis in Section 6.

6 Discussion

Why Sensitivity-VIB is stronger than the latent regularizer. The Jacobian-norm penalty controls $\mathbb{E}_x[\|J_{\mu_\theta}(x)\|_F^2]$, the expected squared Frobenius norm of the encoder Jacobian. This is a stronger and more direct sensitivity control than latent covariance regularization: (i) a small Jacobian norm implies the encoder mean changes little under any input perturbation, including those aligned with leading covariance directions; (ii) a flat Jacobian spectrum discourages the encoder from concentrating sensitivity in a small number of latent directions; and (iii) consistency across clean and perturbed inputs follows directly. This is consistent with Proposition 1, which upper-bounds the shift sensitivity through the encoder Jacobian norm. Empirically, this matches the results in Tables 1 and 2: under CSP2, Sensitivity-VIB leads all methods at strong shift by +72.1 pp on MNIST, +7.8 pp on CIFAR-10, and +11.9 pp on STL-10 over Shrinkage VIB, while under the milder CSP1 both proposed methods remain competitive, with Shrinkage VIB leading on STL-10.

The complementary roles of our two proposed regularizers. Both Sensitivity-VIB and Shrinkage VIB target different aspects of covariance-shift robustness. Shrinkage VIB penalizes the discrepancy between the latent covariance and its isotropic projection, encouraging uniform variance allocation across latent dimensions and improving over Standard VIB across all datasets. Sensitivity-VIB directly penalizes the encoder Jacobian norm, targeting the input-to-latent sensitivity identified by Proposition 1 and Theorem 3 as the dominant source of VIB loss variation under covariance shift. As established in Remark 4, Shrinkage VIB does not bound $\|J_{\mu_\theta}(x)\|_F$, which explains its consistently limited advantage at strong shift under CSP2 relative to Sensitivity-VIB. Nonetheless, Shrinkage VIB achieves the highest mean accuracy on STL-10 under CSP1 (0.592 vs. 0.567), confirming that latent isotropy regularization can be preferable when shift is purely geometric and moderate. The two methods are therefore complementary: Shrinkage VIB offers a lightweight latent-space regularizer with broad gains across shift regimes, while Sensitivity-VIB provides larger margins under compound or severe covariance shift.

Comparison to discriminative Jacobian regularization. Sokolić et al. (21) and Varga et al. (10) use Jacobian penalties to improve adversarial robustness in discriminative classifiers, motivated by margin bounds. Sensitivity-VIB differs in three ways: (i) the penalty is derived from a first-order IB covariance shift bound (Proposition 1), not a margin; (ii) it targets the *encoder* Jacobian J_{μ_θ} (the mean network), not the full classifier Jacobian; and (iii) the weight scaling is motivated by the initialization loss scale and is intended to avoid representation collapse.

Limitation. The local-perturbation limitation applies to the sensitivity-based robustness guarantee: it holds only for bounded perturbations $\|\delta\|_2 \leq \epsilon$ in equation 10 and the first-order bound of Proposition 1. This specific theoretical limitation does not apply to the shrinkage-based formulation.

7 Conclusion

We introduced two complementary VIB loss functions for covariance-shift robustness: Shrinkage VIB, a lightweight latent-space regularizer that encourages isotropic latent covariance via a Ledoit–Wolf-inspired penalty, and Sensitivity-VIB, which augments the VIB objective with a Jacobian-norm penalty motivated by the first-order upper bound on worst-case VIB loss variation under bounded Frobenius covariance perturbations (Proposition 1, Theorem 3). Both methods improve over Standard VIB across all datasets and protocols. Shrinkage VIB provides consistent gains under mild geometric shift, achieving the best mean accuracy on STL-10 under CSP1. Sensitivity-VIB is the stronger method under compound and severe shift, yielding substantially larger margins at $\rho = 1.00$ across all three datasets under CSP2, and the smallest latent covariance drift in most settings (Figures 8, 9, and 10). Evaluation on standard out-of-distribution benchmarks and non-isotropic shift models remains an important direction for future work.

AI use statement

In this work, we used generative AI tools to assist with editing and polishing the writing of some sections. We have not used generative AI tools to generate experimental results, run experiments, or produce the core theoretical derivations in this work. We take responsibility for the final content of this work, including all text, claims, and materials produced with the aid of generative AI.

References

[1] N. Tishby, F. Pereira, and W. Bialek. The information bottleneck method. In *Proc. 37th Annual Allerton Conference on Communication, Control, and Computing*, 1999.

[2] G. Chechik, A. Globerson, N. Tishby, and Y. Weiss. Information bottleneck for Gaussian variables. *Journal of Machine Learning Research*, 6:165–188, 2005.

[3] A. A. Alemi, I. Fischer, J. V. Dillon, and K. Murphy. Deep variational information bottleneck. In *International Conference on Learning Representations (ICLR)*, 2017.

[4] S. Sagawa, P. W. Koh, T. B. Hashimoto, and P. Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *International Conference on Learning Representations (ICLR)*, 2020.

[5] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations (ICLR)*, 2018.

[6] M. Arjovsky, L. Bottou, I. Gulrajani, and D. Lopez-Paz. Invariant risk minimization. *arXiv:1907.02893*, 2019.

[7] B. Sun and K. Saenko. Deep CORAL: Correlation alignment for deep domain adaptation. In *ECCV Workshops*, 2016.

[8] J. Sokolić, R. Giryes, G. Sapiro, and M. R. D. Rodrigues, "Robust large margin deep neural networks," *IEEE Transactions on Signal Processing*, vol. 65, no. 16, pp. 4265–4280, 2017.

[9] J. Hoffman, D. A. Roberts, and S. Yaida, "Robust learning with Jacobian regularization," in *International Conference on Learning Representations (ICLR)*, 2020.

[10] D. Varga, A. Csiszárik, and Z. Zombori. Gradient regularization improves accuracy of discriminative models. *arXiv:1712.09936*, 2018.

[11] S. Rifai, P. Vincent, X. Muller, X. Glorot, and Y. Bengio, "Contractive auto-encoders: Explicit invariance during feature extraction," in *Proc. 28th Int. Conf. Machine Learning (ICML)*, 2011, pp. 833–840. `https://icml.cc/2011/papers/455_icmlpaper.pdf`

[12] O. Ledoit and M. Wolf. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88(2):365–411, 2004.

[13] R. Bhatia. *Matrix Analysis*. Springer, New York, 1997.

[14] C. Davis and W. M. Kahan. The rotation of eigenvectors by a perturbation III. *SIAM Journal on Numerical Analysis*, 7(1):1–46, 1970.

[15] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[16] A. Krizhevsky, "Learning multiple layers of features from tiny images," Technical Report, University of Toronto, 2009.

[17] A. Coates, A. Y. Ng, and H. Lee, "An analysis of single-layer networks in unsupervised feature learning," in *Proc. 14th Int. Conf. Artificial Intelligence and Statistics (AISTATS)*, 2011, pp. 215–223.

[18] I. T. Jolliffe and J. Cadima. Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2065):20150202, 2016.

[19] D. R. Cox. The regression analysis of binary sequences. *Journal of the Royal Statistical Society: Series B (Methodological)*, 20(2):215–242, 1958.

[20] R. A. Vandermeulen, W. M. Tai, and B. Aragam. Dimension-independent rates for structured neural density estimation. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, PMLR 267:60857–60879, 2025.

[21] J. Sokolić, R. Giryes, G. Sapiro, and M. R. D. Rodrigues. Robust large margin deep neural networks. *IEEE Transactions on Signal Processing*, 65(16):4265–4280, 2017.

[22] O. Ledoit and M. Wolf. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88(2):365–411, 2004.

[23] G. Chechik, A. Globerson, N. Tishby, and Y. Weiss. Information bottleneck for Gaussian variables. *Journal of Machine Learning Research*, 6(6):165–188, 2005.

A Relation to Existing Methods and Additional Empirical Diagnostics

Natural images exhibit strong local spatial dependence, with statistical dependence between pixels generally weakening with spatial separation (16); more recent analysis of deep density estimators formalizes this as approximate conditional independence beyond a small Markov neighborhood (20). Consequently, geometric transformations such as rotation and translation do not merely perturb individual pixel values: they reorganize the spatial covariance structure of the pixel grid. This motivates CSP1/CSP2 as a practical, image-native probe of structured covariance perturbation, without claiming that the transformed and original images are themselves correlated in any particular way.

The main text motivates Sensitivity-VIB and Shrinkage-VIB from a first-order covariance-perturbation analysis of the Gaussian IB objective, following the Gaussian IB eigen-structure of Chechik et al. (23) (Section 3). This appendix examines how that motivation relates to existing robustness and covariance-alignment methods — Jacobian regularization (21; 9), CORAL-style covariance alignment (7), distributionally robust optimization (4), and Ledoit–Wolf-style shrinkage estimation (22) — through controlled, matched-protocol comparisons, and reports supplementary diagnostics developed to probe the covariance-shift mechanism itself. It addresses five aspects: (i) whether CSP1/CSP2 isolate covariance change from other distributional effects, and the limits of that claim; (ii) whether encoder sensitivity tracks robustness in the trained nonlinear models, beyond the linear-Gaussian setting of Section 3; (iii) quantitative comparison against JR and CORAL under matched architecture and protocol; (iv) how the proposed objective relates to DRO; and (v) the scope of the covariance-shift claim, including generalization beyond Gaussian latents.

A.1 Covariance-Shift Diagnostics

To characterize what CSP1/CSP2 perturb, we report the normalized covariance displacement

$$\delta_\Sigma(\rho) = \frac{\|\widehat{\Sigma}_{\text{shift}}(\rho) - \widehat{\Sigma}_{\text{clean}}\|_F}{\|\widehat{\Sigma}_{\text{clean}}\|_F},$$

alongside the analogous normalized mean displacement, on CIFAR-10. $\delta_\Sigma(\rho)$ increases from 0 at $\rho = 0$ to 0.497 at $\rho = 1$, confirming that CSP1/CSP2 substantially and monotonically perturb second-order statistics. This diagnostic does not, however, isolate covariance change as the sole cause of the observed robustness gains, since the protocols simultaneously introduce spatial and photometric changes; we return to this limitation below.

For completeness, we record the implementation details underlying this diagnostic and the main experiments: CSP1/CSP2 sample an independent rotation angle for each image and apply the rotation on the discrete pixel grid using bilinear interpolation, with output dimensions fixed to the input dimensions and zero-filling outside the original support. Conditional on a fixed rotation angle θ, the ideal linear rotation satisfies $\Sigma_{X'} = R_\theta \Sigma_X^{(0)} R_\theta^\top$; because CSP1/CSP2 sample θ independently per image, the empirical shifted covariance additionally averages over the rotation-angle distribution, and interpolation and boundary handling introduce further, smaller deviations from this ideal relation, confined mainly to border pixels. For additive independent sensor-style noise ξ, $\Sigma_{X'} = A\Sigma_X^{(0)} A^\top + \Sigma_\xi$. These transformations provide controlled covariance-changing perturbations motivated by the covariance-shift model of Definition 1, but should not be read as exact realizations of it: Definition 1 requires $P_\rho(Y \mid X) = P_0(Y \mid X)$, and label preservation under the transformed pixel-coordinate distribution is not formally established here.

A.2 Controlled Isotropic Covariance Intervention

To probe covariance magnitude in isolation from spatial transformations and eigenspace rotation, we ran a controlled intervention $x' = \mu + \sqrt{1+\rho}\,(x - \mu)$, which gives $\Sigma_X^{(\rho)} = (1+\rho)\Sigma_X^{(0)}$ and hence an exact relative Frobenius displacement of ρ, while leaving all eigenvectors and the eigenvalue ordering of $\Sigma_X^{(0)}$ unchanged. This isolates isotropic covariance scaling from spatial transformations and eigenspace rotation; it does not hold higher-order moments fixed, since these also scale under the transformation.

Table 3: Accuracy under isotropic covariance rescaling (CIFAR-10). This is a negative control that changes covariance magnitude while preserving the eigenspace.

ρ	0	.16	.33	.50	.66	.83	1.00
Std VIB	.792	.791	.789	.785	.784	.777	.769
Shrinkage VIB	.787	.786	.785	.781	.775	.769	.758
Sensitivity VIB	.773	.773	.770	.765	.757	.746	.735

Standard VIB attains slightly higher accuracy than either regularizer under this intervention. This is a negative control, not a validation of the eigenspace mechanism: an isotropic rescaling changes only covariance magnitude, so neither regularizer is expected to have a structural advantage, and the result rules out magnitude change alone as the explanation for the CSP1/CSP2 gains without speaking to eigenspace rotation. For a small rotation $R_\theta \approx I + \theta K$ (K skew-symmetric), $\Delta\Sigma \approx \theta(K\Sigma_X^{(0)} - \Sigma_X^{(0)} K)$ is a commutator, which in general does not share eigenvectors with $\Sigma_X^{(0)}$; this is the regime the Weyl/Davis–Kahan analysis of Section 3 concerns, and it is not instantiated by the isotropic control. We therefore scope the empirical claim to structured covariance perturbations, including anisotropic perturbations that may alter the covariance eigenspace, rather than to covariance shift in general. CSP1/CSP2 are consistent with this mechanism but do not isolate it from other confounded distributional changes; a controlled non-isotropic intervention that rotates the covariance eigenspace directly (e.g., $\Sigma' = A\Sigma A^\top$ for A chosen not to commute with Σ), without image-space geometric transformation, remains an open direction and, together with real-world domain-shift evaluation, is the main empirical gap of this work.

A.3 Encoder Sensitivity in the Trained Nonlinear Models

Proposition 1 and Theorem 3 do not rely on the linear-Gaussian eigen-structure assumption, though Theorem 1's eigen-structure motivation (23) is derived under it. To check whether $\|J_{\mu_\theta}\|_F$ tracks robustness in the trained nonlinear CNN encoders, we ran a λ_{sens} sweep at three regularization strengths (3 seeds each, CIFAR-10).

Across the three regularization strengths, increasing λ_{sens} monotonically reduces the measured Jacobian norm while increasing shifted accuracy and reducing the accuracy drop, with clean accuracy essentially unchanged.

Table 4: Mean Jacobian norm (MJN) vs. robustness across λ_{sens}, averaged over 3 seeds.

λ_{sens}	MJN	Shifted acc. (ρ=1)	Acc. drop
3×10^{-4}	8.50	.530	.210
1×10^{-3}	5.38	.534	.205
3×10^{-3}	3.46	.541	.194

This provides a controlled empirical diagnostic consistent with the sensitivity mechanism, though with only three sweep points it does not constitute a formal nonlinear-theory result, and we do not report a pooled correlation statistic here given the small and heterogeneous sample across method variants.

A.4 Comparison with Jacobian Regularization and Deep CORAL

We implemented JR, which penalizes the full network Jacobian $J_f = J_{\text{clf}} J_{\mu_\theta}$ via a single-projection Hutchinson estimator with λ_{JR} matched in order of magnitude to λ_{sens}, and Deep CORAL, which aligns encoder covariance between clean and shifted batches and requires shifted-domain samples during training — an assumption our methods do not make, since Sensitivity-VIB and Shrinkage-VIB target robustness to *unseen* shift. Both baselines were trained under the same protocol and architecture as the main experiments, on a reduced schedule.

Table 5: Preliminary comparison against JR and CORAL (CIFAR-10, reduced schedule).

ρ	0	.17	.33	.50	.67	.83	1.00	
Std VIB	.755	.616	.586	.576	.563	.544	.522	
Shrinkage VIB	.749	.592	.574	.566	.541	.514	.488	
Sensitivity VIB	.789	.639	.595	.577	.558	.538	.522	
JR	.718	.603	.565	.556	.533	.529	.519	
CORAL	.565	.413	.387	.378	.363	.349	.341	
Acc. drop (clean $\rightarrow$ ρ=1)	Std VIB: .233		Shrinkage: .261		Sensitivity: .267		JR: .199	CORAL: .224

Sensitivity-VIB attains higher absolute accuracy than JR at every tested shift level, but this largely reflects its higher clean accuracy rather than superior robustness: the clean-to-shifted accuracy drop is not smaller for Sensitivity-VIB than for JR in this table (.267 vs. .199). This reduced-schedule, single-architecture comparison therefore does not establish robustness superiority over JR; it shows that penalizing only the encoder mean Jacobian J_{μ_θ} — the quantity Proposition 1 directly bounds for the VIB loss — is competitive with penalizing the full network Jacobian under matched architecture and schedule. CORAL obtains lower accuracy than the VIB variants in this experiment; because CORAL uses shifted-domain samples during training while our methods target unseen shifts without such samples, this reflects a difference in training assumptions and should not be read as a controlled comparison of identical learning settings.

A.5 Relation to Distributionally Robust Optimization

Our formulation admits a DRO interpretation (4) in which uncertainty is placed on the input covariance within a Frobenius-norm ambiguity set (Section 4). We did not include a separate DRO baseline because constructing a representative DRO objective that uses the same covariance uncertainty set, model class, and training information, without introducing additional favorable assumptions, is nontrivial; standard DRO baselines typically require an inner-loop worst-case optimization or a differently shaped uncertainty set, which would complicate a like-for-like comparison. We leave a controlled empirical comparison with covariance-constrained DRO as future work.

A.6 Scope Beyond Gaussian Latents and Image Encoders

The Gaussian assumption is required for the analytically tractable eigen-structure analysis of Section 3 (23) and for the specific Gaussian VIB parameterization $p_\theta(z \mid x) = \mathcal{N}(\mu_\theta(x), \text{diag}(\sigma^2_\theta(x)))$ used in our experiments; the definitions of the two proposed regularizers themselves do not require a Gaussian latent distribution. Shrinkage-VIB operates on the empirical latent covariance and requires only a finite second moment, in the

spirit of Ledoit–Wolf shrinkage estimation (22); Sensitivity-VIB requires only that the encoder mean map be differentiable. The same definitions could in principle be applied to other differentiable encoders, including transformer-based text encoders or world-model latent-state encoders trained with a VIB-style objective; we have not evaluated these settings and do not claim validated performance beyond the image-classification experiments reported here.

We also did not evaluate naturally occurring cross-domain or cross-site distribution shift in this work. The experiments in this paper establish robustness only under the controlled perturbation protocols considered here (CSP1, CSP2, and the isotropic and rotation-based diagnostics above) and should not be read as evidence of real-world domain-shift generalization; validation on naturally occurring covariance shift (e.g., multi-site medical sensing, the motivating application in Section 1) remains future work.

B Optimal Gaussian IB and Effect of covariance perturbations

Eigenvalue shift (Weyl). Let $M^{(\rho)} = M^{(0)} + \Delta M$, and let λ_i and $\lambda_i^{(\rho)}$ denote the corresponding eigenvalues. By applying $\|\Delta M\|_{\mathrm{op}} = \max_{\|x\|_2=1} \|\Delta M x\|_2$, we have Weyl's inequality gives

$$|\lambda_i^{(\rho)} - \lambda_i| \;\le\; \|\Delta M\|_{\mathrm{op}}, \qquad \forall i. \tag{16}$$

Eigenvector perturbation (Davis–Kahan). Let $\mathrm{gap}_i = \min(\lambda_i - \lambda_{i-1},\ \lambda_{i+1} - \lambda_i)$. The Davis–Kahan $\sin\Theta$ theorem yields

$$\sin\angle(v_i, v_i^{(\rho)}) \;\le\; \frac{\|\Delta M\|_2}{\mathrm{gap}_i}. \tag{17}$$

Small eigengaps imply that an encoder subspace can rotate dramatically under a tiny covariance error.

Critical-threshold sensitivity. Substituting $\lambda_i^{(\rho)} = \lambda_i + \delta_i$ into $\beta_i^c \triangleq 1/(1-\lambda_i)$,

$$|\beta_{c,i}^{(\rho)} - \beta_{c,i}| \;\approx\; \frac{|\delta_i|}{(1-\lambda_i)^2}. \tag{18}$$

Therefore, when $\lambda_i \to 1$, even a small perturbation can cause a large change in the critical compression threshold.

Proposition 2 (IB instability near $\lambda = 1$)**.** *Suppose $\|\Delta M\|_2 \le \epsilon$ and $\lambda_i = 1 - \tau$ for small $\tau > 0$. Then $|\beta_{c,i}^{(\rho)} - \beta_{c,i}| \;\le\; \frac{\epsilon}{\tau^2}$ and $\sin\angle(v_i, v_i^{(\rho)}) \;\le\; \frac{\epsilon}{\mathrm{gap}_i}$. Both bounds diverge as $\tau \to 0$ or $\mathrm{gap}_i \to 0$.*

Proof. The bound on $|\beta_{c,i}^{(\rho)} - \beta_{c,i}|$ follows from equation 18 and Weyl's inequality. The angle bound is equation 17. □

One fundamental question is how much the finite-sample covariance estimate under the shifted distribution P_ρ can deviate from that under the nominal distribution P_0. To this end, we establish the following theories to show the gap between the two finite-sample covariance estimates and makes explicit how the severity of the distribution shift and the sample size jointly govern the estimation discrepancy. Formally, this result is captured by the following proposition.

Proposition 3 (Sample covariance deviation under covariance shift — finite-sample bound)**.** *Let $x_1^{(0)}, \dots, x_N^{(0)} \sim P_0$ and $x_1^{(\rho)}, \dots, x_N^{(\rho)} \sim P_\rho$ be i.i.d. zero-mean sub-Gaussian vectors, and define the sample covariances $\widehat{\Sigma}_X^{(r)} = \frac{1}{N}\sum_{i=1}^N x_i^{(r)} (x_i^{(r)})^\top$ for $r \in \{0, \rho\}$. Then for every $\delta \in (0,1)$, there exists a constant $C > 0$ such that with probability at least $1 - \delta$,*

$$\big\|\widehat{\Sigma}_X^{(\rho)} - \widehat{\Sigma}_X^{(0)}\big\|_{\mathrm{op}} \;\le\; \epsilon(\rho) \;+\; C\Big(\|\Sigma_X^{(\rho)}\|_{\mathrm{op}} + \|\Sigma_X^{(0)}\|_{\mathrm{op}}\Big)\left(\sqrt{\frac{d + \log(2/\delta)}{N}} + \frac{d + \log(2/\delta)}{N}\right).$$

Proof. By the triangle inequality applied via adding and subtracting the sample covariances,

$$\widehat{\Sigma}_X^{(\rho)} - \widehat{\Sigma}_X^{(0)} = \big(\widehat{\Sigma}_X^{(\rho)} - \Sigma_X^{(\rho)}\big) + \Delta_\rho + \big(\Sigma_X^{(0)} - \widehat{\Sigma}_X^{(0)}\big),$$

so

$$\left\|\widehat{\Sigma}_X^{(\rho)} - \widehat{\Sigma}_X^{(0)}\right\|_{\mathrm{op}} \;\le\; \left\|\widehat{\Sigma}_X^{(\rho)} - \Sigma_X^{(\rho)}\right\|_{\mathrm{op}} + \|\Delta_\rho\|_{\mathrm{op}} + \left\|\widehat{\Sigma}_X^{(0)} - \Sigma_X^{(0)}\right\|_{\mathrm{op}}.$$

Since $\|\Delta_\rho\|_{\mathrm{op}} \le \|\Delta_\rho\|_F \le \epsilon(\rho)$, it remains to bound the two sample-deviation terms. Applying the Matrix Bernstein inequality to each term independently at confidence level $\delta/2$, there exists $C > 0$ such that with probability at least $1 - \delta/2$ (for each $r \in \{0, \rho\}$),

$$\left\|\widehat{\Sigma}_X^{(r)} - \Sigma_X^{(r)}\right\|_{\mathrm{op}} \;\le\; C\,\|\Sigma_X^{(r)}\|_{\mathrm{op}} \left(\sqrt{\frac{d + \log(2/\delta)}{N}} + \frac{d + \log(2/\delta)}{N}\right).$$

A union bound over $r \in \{0, \rho\}$ ensures both inequalities hold simultaneously with probability at least $1 - \delta$. Substituting and collecting terms yields the stated bound. □

Remark 2. *If one additionally assumes $\|\Sigma_X^{(\rho)}\|_{\mathrm{op}} \le \|\Sigma_X^{(0)}\|_{\mathrm{op}} + \epsilon(\rho)$ (which follows directly from Weyl's inequality), the prefactor simplifies to $2\|\Sigma_X^{(0)}\|_{\mathrm{op}} + \epsilon(\rho)$, making the dependence on the baseline spectrum explicit. In the regime $N \gg d$, the leading term is $O\big(\|\Sigma\|_{\mathrm{op}}\sqrt{d/N}\big)$, showing that the statistical fluctuation is controlled purely by the sample size and intrinsic dimensionality, while $\epsilon(\rho)$ captures the irreducible shift bias.*

Theorem 2 (Mutual information stability under covariance shift)**.** *Let (X, Y) be jointly Gaussian and let $M^{(0)} = \Sigma_{X|Y}(\Sigma_X^{(0)})^{-1}$ with eigenvalues $0 \le \lambda_1 \le \cdots \le \lambda_d < 1$. Suppose $M^{(\rho)} = \Sigma_{X|Y}(\Sigma_X^{(\rho)})^{-1}$ with eigenvalues $\lambda_i^{(\rho)}$. Fix β such that the same m eigen-directions remain active under both $M^{(0)}$ and $M^{(\rho)}$, and define $\tau_{\min} := \min_{i=1,\ldots,m}(1 - \lambda_i)$, $L_0 := \|\Sigma_{X|Y}\|_{\mathrm{op}}\,\|(\Sigma_X^{(0)})^{-1}\|_{\mathrm{op}}^2$. Suppose the smallness condition $\|(\Sigma_X^{(0)})^{-1}\|_{\mathrm{op}}\,\epsilon(\rho) < 1$ holds. Then the Gaussian IB mutual informations $I(Z;Y) = -\frac{1}{2}\sum_{i=1}^m \log \lambda_i(M^{(0)})$, $I^{(\rho)}(Z;Y) = -\frac{1}{2}\sum_{i=1}^m \log \lambda_i(M^{(\rho)})$, satisfy*

$$\left|I^{(\rho)}(Z;Y) - I(Z;Y)\right| \;\le\; \frac{m\,L_0\,\epsilon(\rho)}{\lambda_{\min}(M^{(0)})\left(1 - \|(\Sigma_X^{(0)})^{-1}\|_{\mathrm{op}}\,\epsilon(\rho)\right)}, \tag{19}$$

where $\lambda_{\min}(M^{(0)}) = \min_{i=1,\ldots,m}\lambda_i(M^{(0)})$.

Remark 3. *Three aspects of equation 19 are worth highlighting. First, the bound is fully rigorous: the first-order approximation of Remark 1 is replaced by the exact matrix identity $(B+E)^{-1} - B^{-1} = -B^{-1}E(B+E)^{-1}$, so no higher-order remainder is discarded. Second, the denominator $\lambda_{\min}(M^{(0)})$ reflects the fact that mutual information is most sensitive to perturbations in directions where λ_i is small, i.e., where the encoder has the least redundancy. Third, the bound degrades gracefully as $\epsilon(\rho) \to 0$: the right-hand side of equation 19 vanishes, confirming that no shift implies no instability.*

Proof. The proof proceeds in three steps.

Step 1 (Rigorous M-perturbation bound). Using the exact identity $(B+E)^{-1} - B^{-1} = -B^{-1}E(B+E)^{-1}$ with $B = \Sigma_X^{(0)}$ and $E = \Delta_\rho$, we obtain

$$M^{(\rho)} - M^{(0)} \;=\; -\Sigma_{X|Y}(\Sigma_X^{(0)})^{-1}\Delta_\rho(\Sigma_X^{(\rho)})^{-1}.$$

Taking operator norms and using $\|(\Sigma_X^{(\rho)})^{-1}\|_{\mathrm{op}} \le \|(\Sigma_X^{(0)})^{-1}\|_{\mathrm{op}}/(1 - \|(\Sigma_X^{(0)})^{-1}\|_{\mathrm{op}}\|\Delta_\rho\|_{\mathrm{op}})$, which holds under $\|(\Sigma_X^{(0)})^{-1}\|_{\mathrm{op}}\,\epsilon(\rho) < 1$ since $\|\Delta_\rho\|_{\mathrm{op}} \le \|\Delta_\rho\|_F \le \epsilon(\rho)$,

$$\|\Delta M\|_{\mathrm{op}} \;:=\; \|M^{(\rho)} - M^{(0)}\|_{\mathrm{op}} \;\le\; \frac{L_0\,\epsilon(\rho)}{1 - \|(\Sigma_X^{(0)})^{-1}\|_{\mathrm{op}}\,\epsilon(\rho)} \;=:\; \kappa. \tag{20}$$

Step 2 (Eigenvalue perturbation via Weyl). By Weyl's inequality equation 16,

$$|\lambda_i^{(\rho)} - \lambda_i| \;\le\; \|\Delta M\|_{\mathrm{op}} \;\le\; \kappa, \qquad \forall\, i = 1, \ldots, m.$$

Step 3 (Mutual information difference). Using the mutual information expression and the triangle inequality,

$$\left|I^{(\rho)}(Z;Y) - I(Z;Y)\right| \;=\; \frac{1}{2}\left|\sum_{i=1}^m \log\frac{\lambda_i(M^{(0)})}{\lambda_i(M^{(\rho)})}\right| \;\le\; \frac{1}{2}\sum_{i=1}^m \left|\log\frac{\lambda_i^{(\rho)}}{\lambda_i}\right|. \tag{21}$$

For each term, using $|\log(1+t)| \le |t|/(1-|t|)$ for $|t|<1$ and writing $\lambda_i^{(\rho)} = \lambda_i(1+\delta_i/\lambda_i)$ with $|\delta_i| \le \kappa$,

$$\left|\log \frac{\lambda_i^{(\rho)}}{\lambda_i}\right| \;=\; \left|\log\!\left(1+\frac{\delta_i}{\lambda_i}\right)\right| \;\le\; \frac{\kappa/\lambda_i}{1-\kappa/\lambda_i} \;\le\; \frac{\kappa}{\lambda_i-\kappa} \;\le\; \frac{\kappa}{\lambda_{\min}(M^{(0)})-\kappa},$$

provided $\kappa < \lambda_{\min}(M^{(0)})$. Summing over m active directions and substituting equation 20,

$$\big|I^{(\rho)}(Z;Y) - I(Z;Y)\big| \;\le\; \frac{m\,\kappa}{2\big(\lambda_{\min}(M^{(0)})-\kappa\big)}.$$

In the practically relevant regime where $\kappa \ll \lambda_{\min}(M^{(0)})$, the denominator is bounded below by $\lambda_{\min}(M^{(0)})$, giving the cleaner form equation 19. □

Gradients of the VIB loss with respect to μ and s. Differentiating equation 8 with respect to μ_j and s_j, using $\partial z_j/\partial \mu_j = 1$ and $\partial z_j/\partial s_j = \frac{1}{2}e^{s_j/2}\varepsilon_j$,

$$\frac{\partial \mathcal{L}_{\mathrm{VIB}}}{\partial \mu_j} = \mathbb{E}_\varepsilon\Big[-\nabla_{z_j}\log q_\phi(y \mid z)\Big] + \beta\,\mu_j(x), \tag{22}$$

$$\frac{\partial \mathcal{L}_{\mathrm{VIB}}}{\partial s_j} = \mathbb{E}_\varepsilon\Big[-\tfrac{1}{2}e^{s_j(x)/2}\,\varepsilon_j\,\nabla_{z_j}\log q_\phi(y \mid z)\Big] + \frac{\beta}{2}\big(e^{s_j(x)}-1\big). \tag{23}$$

C Proof of Proposition 1

Proof. The main idea is under the pointwise perturbation equation 10, we show

$$\big|\mathcal{L}_{\mathrm{VIB}}(\theta,\phi;x+\delta,y) - \mathcal{L}_{\mathrm{VIB}}(\theta,\phi;x,y)\big| \;\le\; \epsilon\,\big\|\nabla_x\mathcal{L}_{\mathrm{VIB}}(x,y)\big\|_2 \;+\; \mathcal{O}(\epsilon^2). \tag{24}$$

Expanding through the chain rule equation 11,

$$\big\|\nabla_x\mathcal{L}_{\mathrm{VIB}}(x,y)\big\|_2 \;\le\; \big\|\nabla_\mu\mathcal{L}_{\mathrm{VIB}}\big\|_2\,\|J_\mu(x)\|_{\mathrm{op}} \;+\; \big\|\nabla_s\mathcal{L}_{\mathrm{VIB}}\big\|_2\,\|J_s(x)\|_{\mathrm{op}}, \tag{25}$$

and therefore 12 satisfied.

Step 1: Taylor expansion of the VIB loss under pointwise perturbation.

Since the encoder outputs $\mu_\theta(x)$ and $s_\theta(x)$ are continuously differentiable functions of x, and the VIB loss equation 8 is twice continuously differentiable in (μ, s), the composition $x \mapsto \mathcal{L}_{\mathrm{VIB}}(\theta,\phi;x,y)$ is twice continuously differentiable in x. We additionally assume that the Hessian $\nabla_x^2\mathcal{L}_{\mathrm{VIB}}(\theta,\phi;x,y)$ is bounded in operator norm by a constant $C<\infty$ in a neighborhood of x containing the ball $\{x+\delta : \|\delta\|_2 \le \epsilon\}$. Under this assumption, Taylor's theorem gives

$$\mathcal{L}_{\mathrm{VIB}}(\theta,\phi;x+\delta,y) = \mathcal{L}_{\mathrm{VIB}}(\theta,\phi;x,y) + \langle\nabla_x\mathcal{L}_{\mathrm{VIB}}(\theta,\phi;x,y),\,\delta\rangle + R(\delta), \tag{26}$$

where $\langle A,B\rangle := \mathrm{tr}(A^\top B) = \sum_{i,j}A_{ij}B_{ij}$ for the Frobenius inner product, and $\nabla_x\mathcal{L}_{\mathrm{VIB}}(\theta,\phi;x,y) \in \mathbb{R}^d$ is the gradient of the full VIB loss with respect to the input x, holding the label y and parameters (θ,ϕ) fixed, and $R(\delta)$ is the second-order remainder satisfying

$$|R(\delta)| \;\le\; C\|\delta\|_2^2 \;\le\; C\epsilon^2, \tag{27}$$

by the boundedness of the Hessian and $\|\delta\|_2 \le \epsilon$.

Step 2: Bounding the loss difference via Cauchy–Schwarz.

Rearranging equation 26,

$$\mathcal{L}_{\mathrm{VIB}}(\theta,\phi;x+\delta,y) - \mathcal{L}_{\mathrm{VIB}}(\theta,\phi;x,y) = \langle\nabla_x\mathcal{L}_{\mathrm{VIB}}(x,y),\,\delta\rangle + R(\delta). \tag{28}$$

Taking absolute values of both sides and applying the triangle inequality,

$$\big|\mathcal{L}_{\mathrm{VIB}}(\theta,\phi;x+\delta,y) - \mathcal{L}_{\mathrm{VIB}}(\theta,\phi;x,y)\big| \le \big|\langle\nabla_x\mathcal{L}_{\mathrm{VIB}}(x,y),\,\delta\rangle\big| + |R(\delta)|. \tag{29}$$

Applying the Cauchy–Schwarz inequality to the inner product term,

$$\big|\langle\nabla_x\mathcal{L}_{\mathrm{VIB}}(x,y),\,\delta\rangle\big| \le \big\|\nabla_x\mathcal{L}_{\mathrm{VIB}}(x,y)\big\|_2\,\|\delta\|_2 \;\le\; \epsilon\,\big\|\nabla_x\mathcal{L}_{\mathrm{VIB}}(x,y)\big\|_2, \tag{30}$$

where the second inequality uses $\|\delta\|_2 \le \epsilon$. Substituting equation 30 and equation 27 into equation 29,

$$\left|\mathcal{L}_{\text{VIB}}(\theta,\phi;x+\delta,y) - \mathcal{L}_{\text{VIB}}(\theta,\phi;x,y)\right| \le \epsilon \left\|\nabla_x \mathcal{L}_{\text{VIB}}(x,y)\right\|_2 + C\epsilon^2, \tag{31}$$

which is equation 24 since $C\epsilon^2 = \mathcal{O}(\epsilon^2)$.

Step 3: Chain-rule decomposition of $\nabla_x \mathcal{L}_{\text{VIB}}$.

The VIB loss equation 8 depends on x only through the encoder outputs $\mu_\theta(x) \in \mathbb{R}^m$ and $s_\theta(x) \in \mathbb{R}^m$. Applying the chain rule,

$$\nabla_x \mathcal{L}_{\text{VIB}}(\theta,\phi;x,y) = J_\mu(x)^\top \nabla_\mu \mathcal{L}_{\text{VIB}}(\theta,\phi;x,y) + J_s(x)^\top \nabla_s \mathcal{L}_{\text{VIB}}(\theta,\phi;x,y), \tag{32}$$

where $J_\mu(x) = \nabla_x \mu_\theta(x) \in \mathbb{R}^{m\times d}$ and $J_s(x) = \nabla_x s_\theta(x) \in \mathbb{R}^{m\times d}$ are the Jacobians of the encoder mean and log-variance with respect to the input, and $\nabla_\mu \mathcal{L}_{\text{VIB}} \in \mathbb{R}^m$, $\nabla_s \mathcal{L}_{\text{VIB}} \in \mathbb{R}^m$ are the gradients of the VIB loss with respect to μ and s, given explicitly by equation 22–equation 23.

Taking the Euclidean norm of both sides of equation 32 and applying the triangle inequality,

$$\left\|\nabla_x \mathcal{L}_{\text{VIB}}(x,y)\right\|_2 \le \left\|J_\mu(x)^\top \nabla_\mu \mathcal{L}_{\text{VIB}}\right\|_2 + \left\|J_s(x)^\top \nabla_s \mathcal{L}_{\text{VIB}}\right\|_2. \tag{33}$$

For any matrix $J \in \mathbb{R}^{m\times d}$ and vector $g \in \mathbb{R}^m$, the definition of the operator norm gives

$$\|J^\top g\|_2 = \sup_{\|v\|_2=1} \langle J^\top g,\, v\rangle = \sup_{\|v\|_2=1} \langle g,\, Jv\rangle \;\le\; \|g\|_2 \sup_{\|v\|_2=1} \|Jv\|_2 = \|g\|_2 \,\|J\|_{\text{op}}. \tag{34}$$

Applying equation 34 to each term in equation 33 with $(J,g) = (J_\mu(x), \nabla_\mu \mathcal{L}_{\text{VIB}})$ and $(J,g) = (J_s(x), \nabla_s \mathcal{L}_{\text{VIB}})$ respectively,

$$\left\|\nabla_x \mathcal{L}_{\text{VIB}}(x,y)\right\|_2 \le \left\|\nabla_\mu \mathcal{L}_{\text{VIB}}\right\|_2 \|J_\mu(x)\|_{\text{op}} + \left\|\nabla_s \mathcal{L}_{\text{VIB}}\right\|_2 \|J_s(x)\|_{\text{op}}, \tag{35}$$

which is equation 25.

Step 4: Worst-case upper bound via supremum over δ.

From equation 26 and the triangle inequality, for any δ with $\|\delta\|_2 \le \epsilon$,

$$\mathcal{L}_{\text{VIB}}(\theta,\phi;x+\delta,y) \le \mathcal{L}_{\text{VIB}}(\theta,\phi;x,y) + \left|\langle \nabla_x \mathcal{L}_{\text{VIB}}(x,y), \delta\rangle\right| + C\epsilon^2. \tag{36}$$

The inner product $\langle \nabla_x \mathcal{L}_{\text{VIB}}(x,y), \delta\rangle$ is maximized over $\|\delta\|_2 \le \epsilon$ by choosing

$$\delta^* \;=\; \frac{\epsilon\, \nabla_x \mathcal{L}_{\text{VIB}}(x,y)}{\|\nabla_x \mathcal{L}_{\text{VIB}}(x,y)\|_2}, \tag{37}$$

the direction of steepest ascent, which satisfies $\|\delta^*\|_2 = \epsilon$ and gives

$$\sup_{\|\delta\|_2\le\epsilon} \langle \nabla_x \mathcal{L}_{\text{VIB}}(x,y), \delta\rangle = \epsilon \left\|\nabla_x \mathcal{L}_{\text{VIB}}(x,y)\right\|_2. \tag{38}$$

Taking the supremum over all $\|\delta\|_2 \le \epsilon$ in equation 36, substituting equation 38, and then applying equation 35,

$$\begin{aligned}\sup_{\|\delta\|_2\le\epsilon} \mathcal{L}_{\text{VIB}}(\theta,\phi;x+\delta,y) &\le \mathcal{L}_{\text{VIB}}(\theta,\phi;x,y) + \epsilon \left\|\nabla_x \mathcal{L}_{\text{VIB}}(x,y)\right\|_2 + \mathcal{O}(\epsilon^2) \\ &\le \mathcal{L}_{\text{VIB}}(\theta,\phi;x,y) + \epsilon\Big(\left\|\nabla_\mu \mathcal{L}_{\text{VIB}}\right\|_2 \|J_\mu(x)\|_{\text{op}} + \left\|\nabla_s \mathcal{L}_{\text{VIB}}\right\|_2 \|J_s(x)\|_{\text{op}}\Big) + \mathcal{O}(\epsilon^2),\end{aligned}$$

which is equation 12. This completes the proof. □

D Theorem 3

Theorem 3 (Jacobian proxy for encoder sensitivity)**.** *Let $x \in \mathbb{R}^d$ and let $\varepsilon \sim \mathcal{N}(0, I_d)$. Assume that the encoder mean $\mu_\theta : \mathbb{R}^d \to \mathbb{R}^m$ is twice continuously differentiable, and let $J_\mu(x) = \nabla_x \mu_\theta(x) \in \mathbb{R}^{m\times d}$ denote its Jacobian. Then for any $\epsilon_0 > 0$,*

$$\mathbb{E}_\varepsilon\left[\frac{\left\|\mu_\theta(x+\epsilon_0\varepsilon) - \mu_\theta(x)\right\|_2^2}{\epsilon_0^2}\right] \;=\; \|J_\mu(x)\|_F^2 \;+\; \mathcal{O}(\epsilon_0^2), \tag{39}$$

so that the finite-difference estimator

$$\widehat{\mathcal{S}}(x;\theta) \;:=\; \frac{\left\|\mu_\theta(x+\epsilon_0\,\varepsilon) - \mu_\theta(x)\right\|_2^2}{\epsilon_0^2}, \qquad \varepsilon \sim \mathcal{N}(0, I_d), \tag{40}$$

is an unbiased estimator of $\|J_\mu(x)\|_F^2$ up to second-order corrections in ϵ_0.

Proof. Apply a first-order Taylor expansion of μ_θ at x:

$$\mu_\theta(x+\epsilon_0\varepsilon) \;=\; \mu_\theta(x) + J_\mu(x)\,\epsilon_0\varepsilon + R(x,\epsilon_0\varepsilon),$$

where the remainder satisfies $\|R(x,\epsilon_0\varepsilon)\|_2 = \mathcal{O}(\epsilon_0^2\|\varepsilon\|_2^2)$ by the twice continuous differentiability of μ_θ. Subtracting $\mu_\theta(x)$ and dividing by ϵ_0,

$$\frac{\mu_\theta(x+\epsilon_0\varepsilon)-\mu_\theta(x)}{\epsilon_0} \;=\; J_\mu(x)\,\varepsilon + \frac{R(x,\epsilon_0\varepsilon)}{\epsilon_0}.$$

Taking the squared ℓ_2 norm and expectation over $\varepsilon\sim\mathcal{N}(0,I_d)$,

$$\mathbb{E}_\varepsilon\left[\frac{\|\mu_\theta(x+\epsilon_0\varepsilon)-\mu_\theta(x)\|_2^2}{\epsilon_0^2}\right] = \mathbb{E}_\varepsilon\big[\|J_\mu(x)\varepsilon\|_2^2\big] + \mathcal{O}(\epsilon_0^2), \tag{41}$$

where the cross term $2\mathbb{E}_\varepsilon[\langle J_\mu(x)\varepsilon, R(x,\epsilon_0\varepsilon)/\epsilon_0\rangle]$ vanishes at leading order since ε is zero-mean and R is second-order in $\epsilon_0\varepsilon$. For the leading term, since $\varepsilon\sim\mathcal{N}(0,I_d)$,

$$\begin{aligned}\mathbb{E}_\varepsilon\big[\|J_\mu(x)\varepsilon\|_2^2\big] &= \mathbb{E}_\varepsilon\big[\mathrm{tr}\big(J_\mu(x)\varepsilon\varepsilon^\top J_\mu(x)^\top\big)\big] = \mathrm{tr}\big(J_\mu(x)\,\mathbb{E}[\varepsilon\varepsilon^\top]\,J_\mu(x)^\top\big)\\ &= \mathrm{tr}\big(J_\mu(x)\,I_d\,J_\mu(x)^\top\big) = \mathrm{tr}\big(J_\mu(x)^\top J_\mu(x)\big) = \|J_\mu(x)\|_F^2.\end{aligned} \tag{42}$$

Substituting equation 42 into equation 41 yields equation 39, completing the proof. □

Theorem 3 confirms that $\widehat{\mathcal{S}}(x;\theta)$ is a consistent single-sample estimator of $\|J_\mu(x)\|_F^2$, which appears in the dominant term of the chain-rule bound equation 25 via $\|J_\mu(x)\|_{\mathrm{op}} \le \|J_\mu(x)\|_F$. The J_s term in equation 25 is omitted for tractability; it is second-order for small β and contributes negligibly in the regime $\beta \ll 1$ used in our experiments.

E The mathematical foundations of Shrinkage VIB

Latent Covariance on Shrinkage VIB. Shrinkage VIB adds a latent-space geometry penalty that encourages the encoder outputs to have approximately isotropic covariance. Given a minibatch $\{x_i\}_{i=1}^B$ and encoder means $\mu_i = \mu_\theta(x_i)\in\mathbb{R}^m$, define the empirical latent covariance $C(\mu) = \frac{1}{B}\sum_{i=1}^B(\mu_i-\bar\mu)(\mu_i-\bar\mu)^\top, \qquad \bar\mu = \frac{1}{B}\sum_{i=1}^B\mu_i$. We then regularize the deviation from the isotropic covariance target with equation 14. This term does not act on the input space directly; rather, it shapes the *geometry of the latent representation* by discouraging anisotropic variance allocation across coordinates.

Lemma 1 (Isotropic Projection and Eigenvalue-Dispersion Form)**.** *For any symmetric matrix $C\succeq 0$, the matrix $\Pi_{\mathrm{iso}}(C) := \frac{\mathrm{tr}(C)}{m}\mathbf{I}$ is the unique minimizer of $\min_{a\in\mathbb{R}}\|C-a\mathbf{I}\|_F^2$. Equivalently, $\mathcal{R}_{\mathrm{shrink}}(\mu)$ is the squared distance from $C(\mu)$ to the closest isotropic covariance with the same trace. Moreover, letting $\lambda_1,\dots,\lambda_m$ be the eigenvalues of $C(\mu)$, $\mathcal{R}_{\mathrm{shrink}}(\mu) = \sum_{j=1}^m\big(\lambda_j-\bar\lambda\big)^2$, $\bar\lambda = \frac{1}{m}\sum_{j=1}^m\lambda_j$. Hence, $\mathcal{R}_{\mathrm{shrink}}(\mu)$ penalizes eigenvalue dispersion and is zero if and only if the latent covariance is isotropic.*

Proof. **(Isotropic projection.)** The objective expands as

$$\|C-aI\|_F^2 = \|C\|_F^2 - 2a\,\mathrm{tr}(C) + ma^2.$$

Differentiating with respect to a and setting the derivative to zero yields $a^\star = \mathrm{tr}(C)/m$. The strict convexity in a gives uniqueness.

(Eigenvalue-dispersion form.) Since $C(\mu)$ is symmetric, write $C = U\Lambda U^\top$ with $\Lambda = \mathrm{diag}(\lambda_1,\dots,\lambda_m)$. The Frobenius norm is orthogonally invariant, so

$$\left\|C-\frac{\mathrm{tr}(C)}{m}I\right\|_F^2 = \big\|\Lambda-\bar\lambda I\big\|_F^2 = \sum_{j=1}^m(\lambda_j-\bar\lambda)^2.$$

The quantity vanishes exactly when all eigenvalues are equal. □

Remark 4 (What shrinkage does and does not control)**.** *The penalty in equation 14 controls the* shape *of the latent covariance but does not directly constrain the* input-output sensitivity *of the encoder. Since $\mathcal{R}_{\mathrm{shrink}}(\mu)$ is a function of the minibatch covariance $C(\mu)$ only, it encourages isotropy of the output distribution on the current batch but does not act on derivatives of μ_θ with respect to the input. In particular, minimizing $\mathcal{R}_{\mathrm{shrink}}(\mu)$ alone does not provide an explicit bound on the Jacobian norm $\|J_{\mu_\theta}(x)\|_F$ or on the response of the latent representation under input-space covariance perturbations. Covariance isotropization in latent space can therefore be achieved without a corresponding guarantee of robustness to perturbations in the input space.*

Although Shrinkage VIB improves latent isotropy, its effect is confined to the geometry of the learned codes. To obtain explicit control over sensitivity to input-space covariance perturbations, we next introduce a Jacobian-based regularizer that directly penalizes encoder variability under structured perturbations.

Sensitivity-VIB, by contrast, acts at the *input-encoder interface*: it directly penalizes the Jacobian norm $\mathbb{E}_x[\|J_{\mu_\theta}(x)\|_F^2]$, which upper-bounds the worst-case change in the VIB loss under any bounded covariance perturbation (Theorem **??**). This is a strictly stronger form of regularization: a small Jacobian norm implies a nearly isotropic latent distribution under mild conditions, but not vice versa. Our ablation (Table **??**) confirms this empirically: Shrinkage VIB achieves 0.672 mean accuracy on MNIST vs. **0.815** for Sensitivity-VIB, a gap of over 14 percentage points.

F Covariance-shift protocols, Covariance Stability, Hyperparameters and architectures

Covariance shift protocol-1 (CSP1). In the first covariance shift protocol, at test time, we apply a parameterized geometric perturbation at shift strength $\rho \in [0, 1]$: rotation $\theta \sim \text{Uniform}(-20\rho, +20\rho)$ degrees and integer translation $(t_x, t_y) \sim \text{round}(\text{Uniform}(-6\rho, +6\rho))$ pixels, applied independently per image via bilinear interpolation. At $\rho = 0$ the transform is the identity; at $\rho = 1$ it produces rotations up to ± 20ř and translations up to ± 6 pixels. No additive noise or contrast scaling is applied, so the dominant effect is a pure geometric covariance shift consistent with the theory.

Covariance Shift Protocol-2 (CSP2). At test time only, each image independently undergoes geometric and photometric perturbations scaled by $\rho \in [0, 1]$, intended to alter input feature covariances and thereby motivate the regularizers evaluated here. Each image is rotated by $\theta \sim \text{Uniform}(-20\rho, 20\rho)$ degrees and translated by $t_x, t_y \sim \text{round}(\text{Uniform}(-3\rho, 3\rho))$ pixels via bilinear interpolation, then rescaled by contrast factor $c \sim \text{Uniform}(1 - 0.2\rho, 1 + 0.2\rho)$ and corrupted by pixelwise noise $\varepsilon \sim \mathcal{N}(0, (0.2\rho)^2)$. RGB images additionally receive per-channel multiplicative scaling $\delta_k \sim \mathcal{N}(1, (0.05\rho)^2)$. Pixel values are clipped to $[0, 1]$.

Latent Covariance Stability: Computation. For each VIB variant, the encoder mean map $\mu_\phi(\cdot)$ (evaluated deterministically, without sampling) is applied to the training split and the shifted test split to obtain latent means $\{\mu_\phi(x_i)\}_{i=1}^N \subset \mathbb{R}^d$, with $d = 16$. The empirical covariance is then estimated as $\widehat{C} = \frac{1}{N-1}\sum_{i=1}^N (\mu_\phi(x_i) - \bar{\mu})(\mu_\phi(x_i) - \bar{\mu})^\top$. The training covariance $\widehat{C}_{\text{train}}$ is computed once on the training split. For each ρ, a fresh shifted test split is generated using an independent seed, and $\widehat{C}_{\text{test}}(\rho)$ is computed in the same way. The reported metric is $\Delta_\mu(\rho) = \|\widehat{C}_{\text{train}} - \widehat{C}_{\text{test}}(\rho)\|_F$, computed.

F.1 Neural Network Architectures

Overview. All VIB variants (Methods 3–5) share *identical* PyTorch `nn.Module` encoder and classifier architectures; only the training loss differs. The Raw CNN and PCA baselines are architecturally independent.

(1) Raw CNN Baseline

The Raw CNN baseline is a shallow convolutional classifier used as a non-VIB reference model. It consists of three convolutional layers with 32, 64, and 128 filters, respectively, where the first two layers are followed by batch normalization, ReLU, and max pooling, and the final convolution is followed by ReLU and global average pooling. The pooled features are passed to a fully connected layer that outputs 10 logits, and the predicted class is obtained by taking the argmax of these logits. The model is trained with cross-entropy loss for 50 epochs using Adam.

(2) PCA + One-vs-Rest Logistic Regression Baseline

Training images are flattened to $X_{\text{tr}}^{\text{flat}} \in \mathbb{R}^{N \times d}$. PCA is applied on the centered matrix, retaining $k = \min(50, \text{rank}(X_{\text{tr}}^{\text{flat}}))$ principal components. Each of the ten binary classifiers in the `OneVsRestClassifier` wraps an identical `Pipeline(StandardScaler, LogisticRegression)` with L2 regularization (`penalty='l2'`), the `lbfgs` solver, and `max_iter=1000`; the scaler is fit independently per binary problem. The predicted class is $\hat{c} = \text{argmax}_j\ s_j$, where s_j is the decision score for class j.

Table 6: Hyperparameters used in all experiments.

Category	Hyperparameter	Value
Architecture	Latent dimension d	16
	Classifier hidden dimension	64
	Input channels	1 (MNIST), 3 (CIFAR-10)
Optimization	Optimizer	Adam
	Learning rate	10^{-3}
	Mini-batch size	128
	Number of epochs	50
	Raw CNN pre-training epochs	50
VIB Regularization	KL penalty β	10^{-3}
	Shrinkage penalty λ_{shrink} (Best)	0.05
	Sensitivity penalty λ_{sens} (Best)	10^{-3}
PCA	PCA components	$\min(50, \text{rank}(X_{\text{train}}))$
	Classifier	One-vs-Rest ℓ_2 logistic regression
Shift Evaluation	Shift grid	`np.linspace(0, 1, 7)`
	Accuracy eval. seed	$1000 + i$
	Covariance eval. seed	$3000 + i$
Reproducibility	Global random seed	42
	Sensitivity perturbation ε_J	10^{-2}

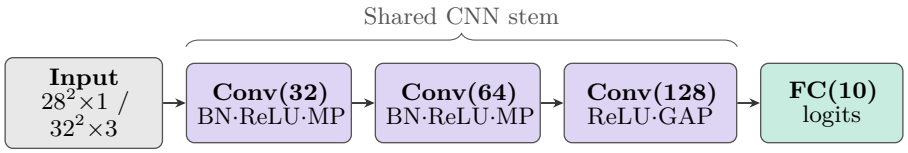


Figure 2: Raw CNN baseline (`RawCNN`). All convolutions use 3×3 filters with `same` padding. BN = Batch Normalization; MP = MaxPool (2×2, stride 2); GAP = Global Average Pooling; FC = Fully Connected. The network returns raw logits; softmax is applied implicitly inside `F.cross_entropy` during training and is not part of the module. Trained for 50 epochs, Adam, lr $= 10^{-3}$, batch size 128. The output is a logit vector $\ell \in \mathbb{R}^{10}$; the predicted class is $\hat{c} = \text{argmax}_j\ \ell_j$.

(3–5) Shared VIB Encoder and Classifier

The `Encoder` module applies the shared convolutional stem and ends with FC($2d_z$):

$$f_\theta(x) = \big[\mu(x),\ \log\sigma^2(x)\big]\ \in\ \mathbb{R}^{2d_z}, \quad d_z = 16. \tag{43}$$

The stochastic latent variable is sampled via the reparameterization trick:

$$z = \mu(x) + \exp\Big(\tfrac{1}{2}\log\sigma^2(x)\Big) \odot \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, I_{d_z}). \tag{44}$$

The `Classifier` module is a two-layer MLP returning logits:

$$\ell_\phi(z) = W_2\ \text{ReLU}(W_1 z + b_1) + b_2, \qquad W_1 \in \mathbb{R}^{64\times16},\ W_2 \in \mathbb{R}^{10\times64}. \tag{45}$$

The Raw CNN is trained with `F.cross_entropy`, which combines log-softmax and negative log-likelihood internally. In the VIB loss, the classifier outputs logits, which are converted to probabilities by an explicit `F.softmax` inside `vib_loss` before computing the cross-entropy term $-\mathbb{E}\big[\sum_j t_j \log(p_j + \epsilon)\big]$, where t_j is the one-hot target and $\epsilon = 10^{-8}$ is a numerical stability constant. At inference, `eval_vib` passes only $\mu(x)$ to the classifier (no sampling noise); the predicted class is $\hat{c} = \text{argmax}_j\ \ell_{\phi,j}(\mu(x))$.

G Per-Dataset Accuracy and Covariance Stability Plots

The following figures show test accuracy and latent covariance stability for each dataset separately, combining CSP1 (top) and CSP2 (bottom) in each panel for easier per-dataset comparison.

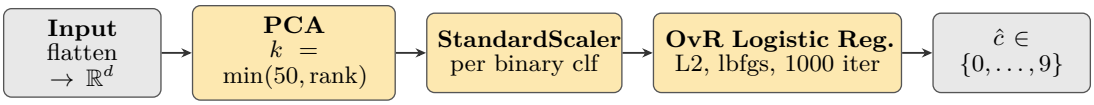


Figure 3: PCA + one-vs-rest logistic regression baseline. No neural network. Features are k-dimensional PCA projections of flattened pixels ($d = 784$ for MNIST, $d = 3{,}072$ for CIFAR-10). `StandardScaler` and `LogisticRegression` are wrapped in a `Pipeline` that is itself wrapped by `OneVsRestClassifier`, so each binary classifier fits its own scaler independently.

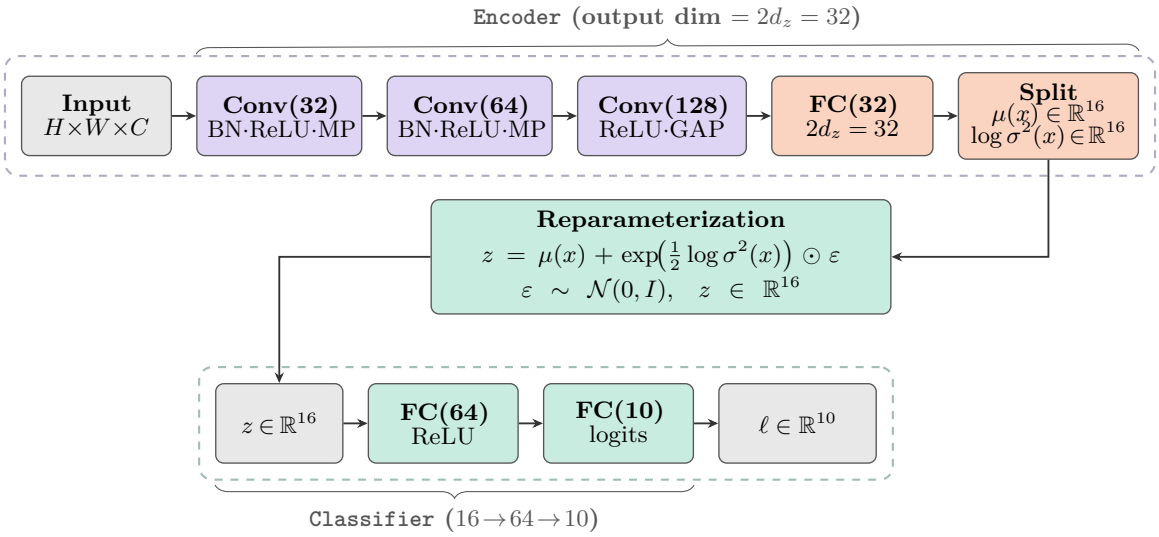
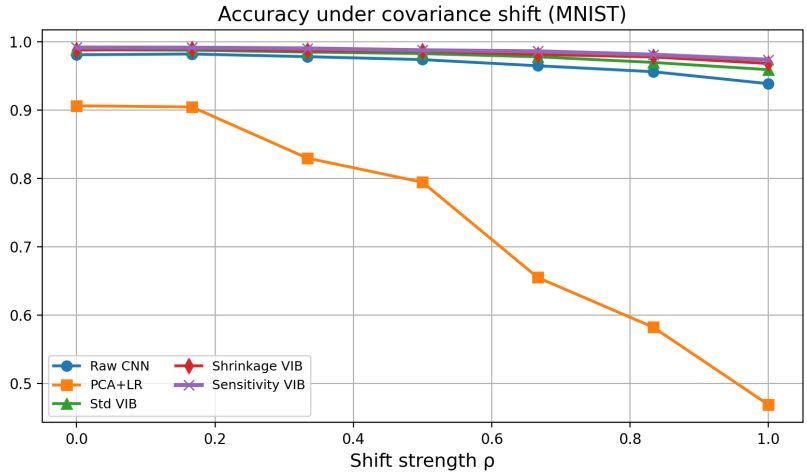


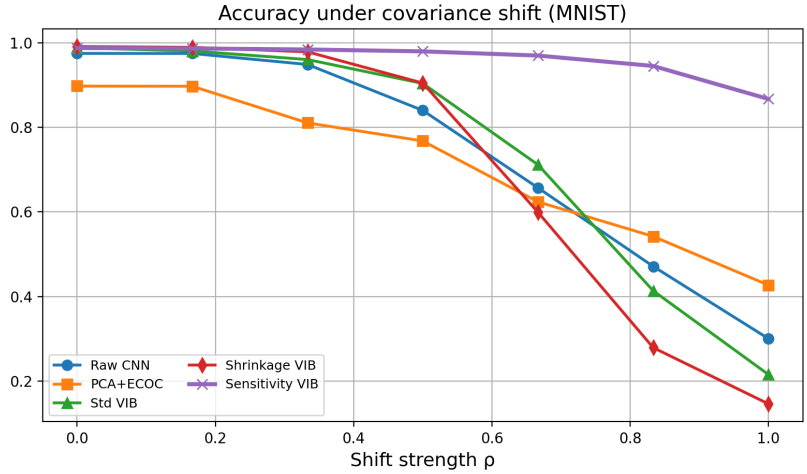
Figure 4: Shared `Encoder` and `Classifier` modules used by all VIB variants (Methods 3–5), implemented as PyTorch `nn.Module` subclasses and composed inside `VIBModel`. Latent dimension $d_z = 16$; the encoder outputs $2d_z = 32$ values, split into mean $\mu(x)$ and log-variance $\log \sigma^2(x)$. The classifier returns a logit vector $\ell \in \mathbb{R}^{10}$; softmax is applied explicitly inside `vib_loss` during training and is not part of the module itself. BN = Batch Normalization; MP = MaxPool (2×2, stride 2); GAP = Global Average Pooling; FC = Fully Connected. Trained for 50 epochs, Adam, lr $= 10^{-3}$, batch size 128.

H Test images across different datasets

This section shows representative training and test images from MNIST, CIFAR-10, and STL-10 under CSP1 and CSP2. The figures provide a visual comparison of the datasets and show how the test images change under the two shift settings, with CSP2 producing the stronger perturbation.

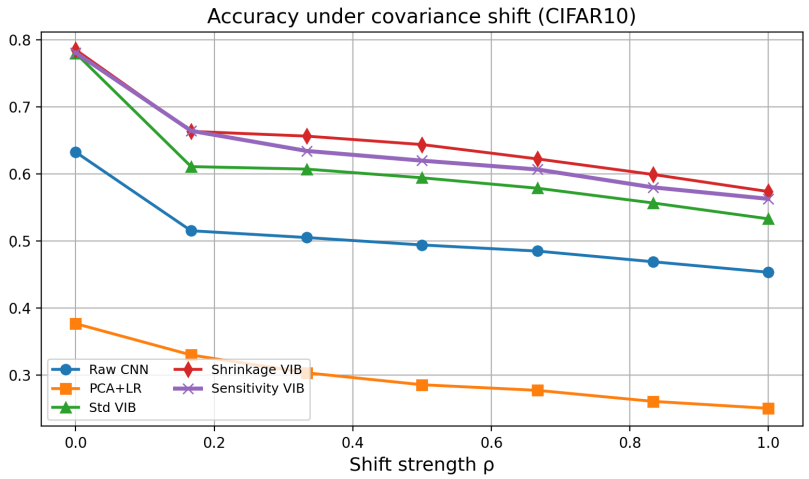


(a) Test Accuracy, in %, vs Shift Strength for MNIST Dataset in CSP1

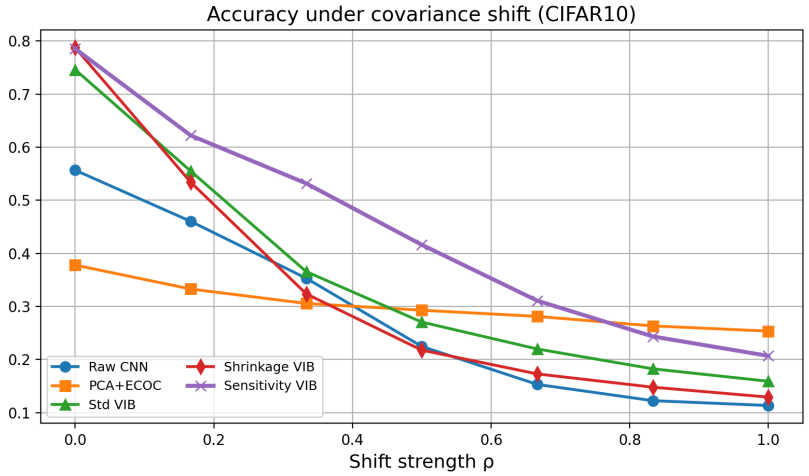


(b) Test Accuracy, in %, vs Shift Strength for MNIST Dataset in CSP2

Figure 5: Test accuracy, in percentage, vs. shift strength ρ for all methods on MNIST. Top: CSP1 (pure geometric shift). Bottom: CSP2 (CP1+geometric and photometric shift).

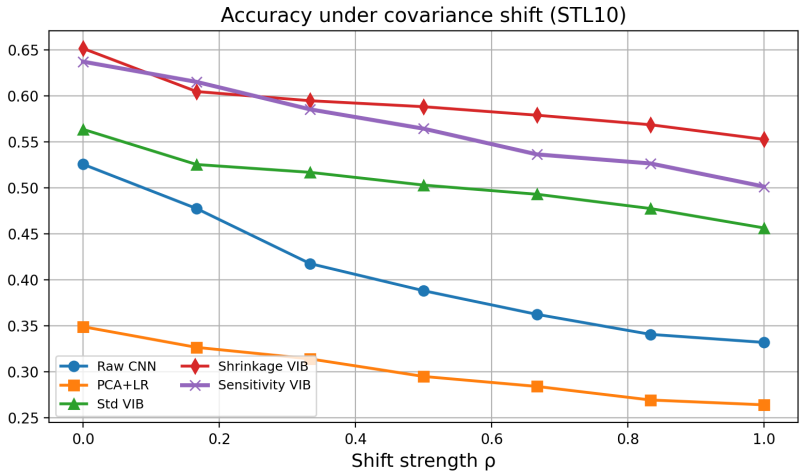


(a) Test Accuracy, in %, vs Shift Strength for CIFAR-10 Dataset in CSP1

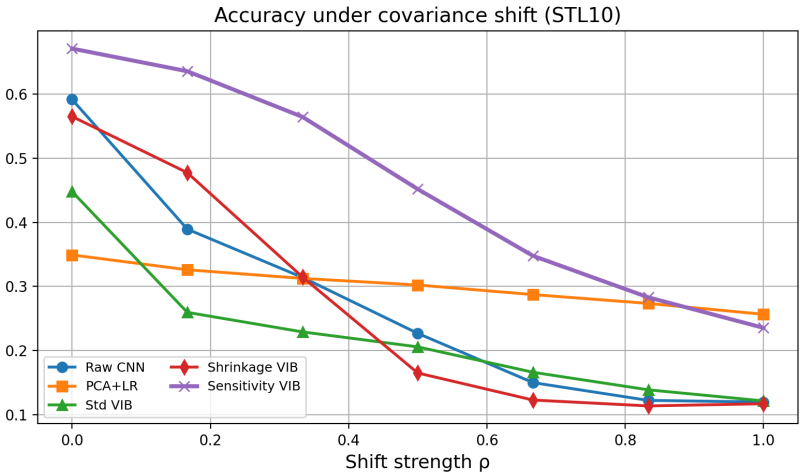


(b) Test Accuracy, in %, vs Shift Strength for CIFAR-10 Dataset in CSP2

Figure 6: Test accuracy, in percentage, vs. shift strength ρ for all methods on CIFAR-10. Top: CSP1 (pure geometric shift). Bottom: CSP2 (CP1+geometric and photometric shift).

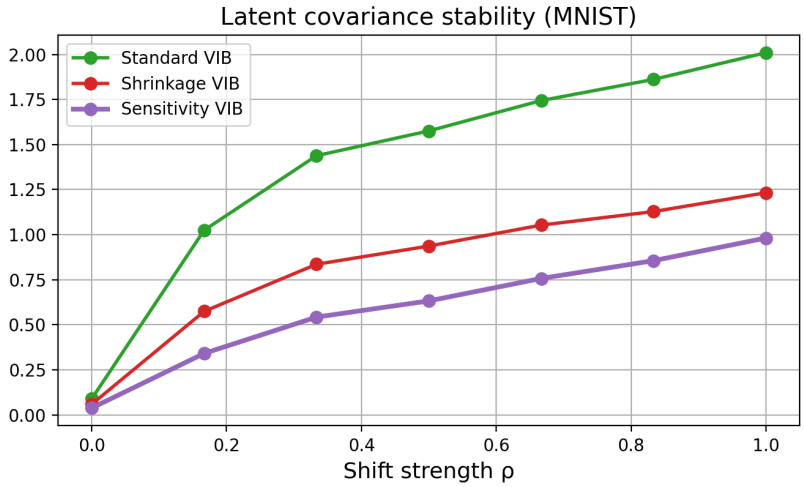


(a) Test Accuracy, in %, vs Shift Strength for STL-10 Dataset in CSP1

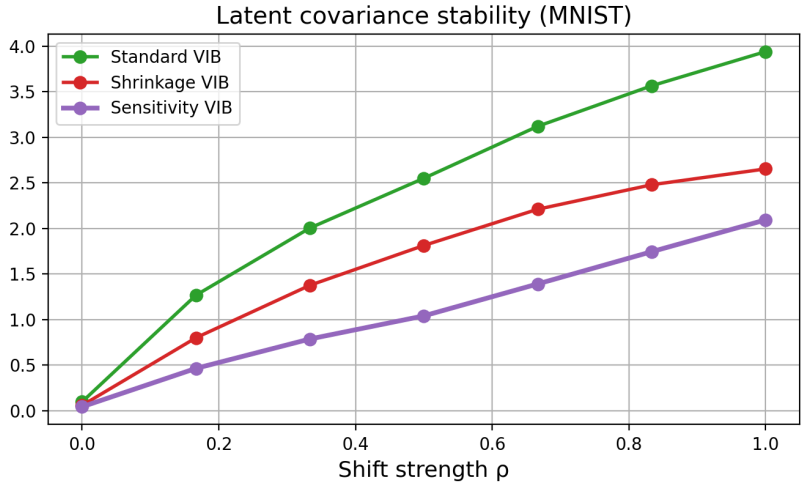


(b) Test Accuracy, in %, vs Shift Strength for STL-10 Dataset in CSP2

Figure 7: Test accuracy, in percentage, vs. shift strength ρ for all methods on STL-10. Top: CSP1 (pure geometric shift). Bottom: CSP2 (CP1+geometric and photometric shift).

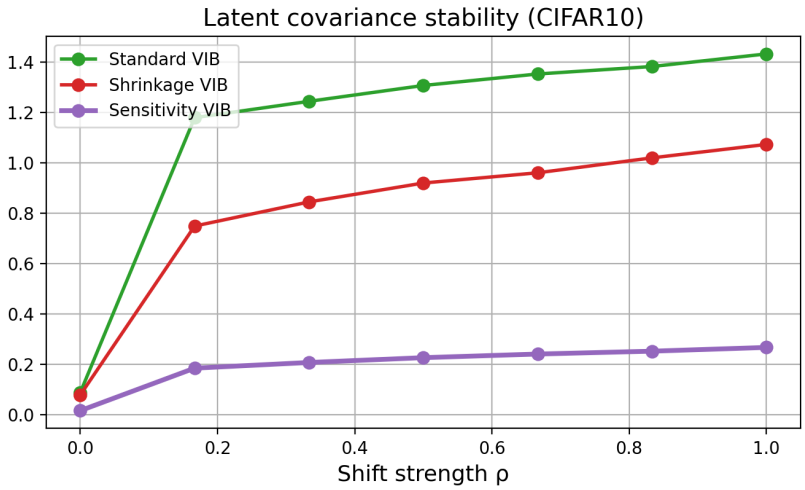


(a) Covariance Stability vs Shift Strength for MNIST Dataset in CSP1

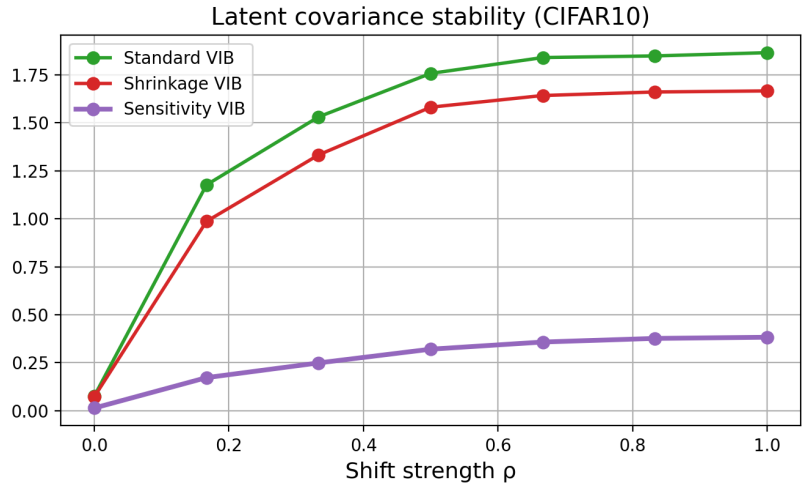


(b) Covariance Stability vs Shift Strength for MNIST Dataset in CSP2

Figure 8: Latent covariance stability, measured by the Frobenius distance $\Delta_\mu(\rho) = \|\widehat{C}_{\text{train}} - \widehat{C}_{\text{test}}(\rho)\|_F$ vs. shift strength ρ for all VIB variants on MNIST. Top: CSP1. Bottom: CSP2.

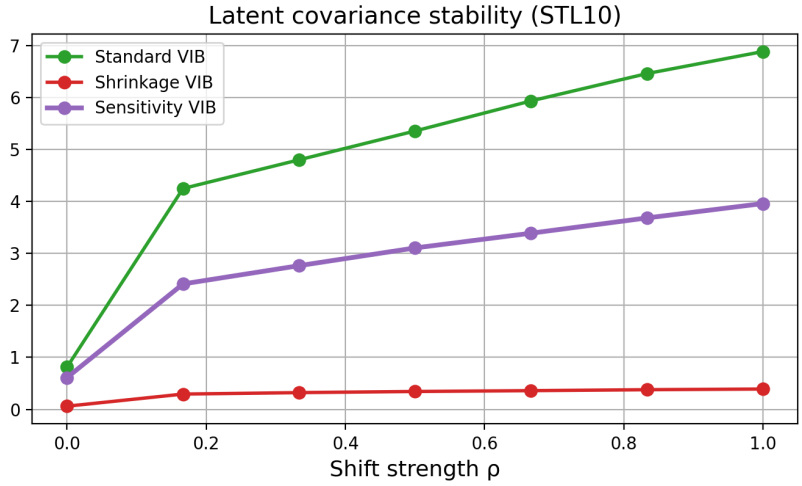


(a) Covariance Stability vs Shift Strength for CIFAR-10 Dataset in CSP1

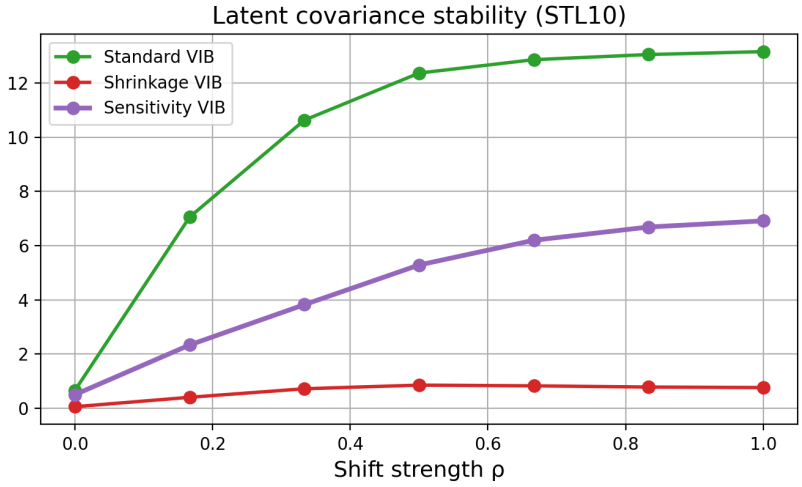


(b) Covariance Stability vs Shift Strength for CIFAR-10 Dataset in CSP2

Figure 9: Latent covariance stability, measured by the Frobenius distance $\Delta_\mu(\rho) = \|\widehat{C}_{\text{train}} - \widehat{C}_{\text{test}}(\rho)\|_F$ vs. shift strength ρ for all VIB variants on CIFAR-10. Top: CSP1. Bottom: CSP2.

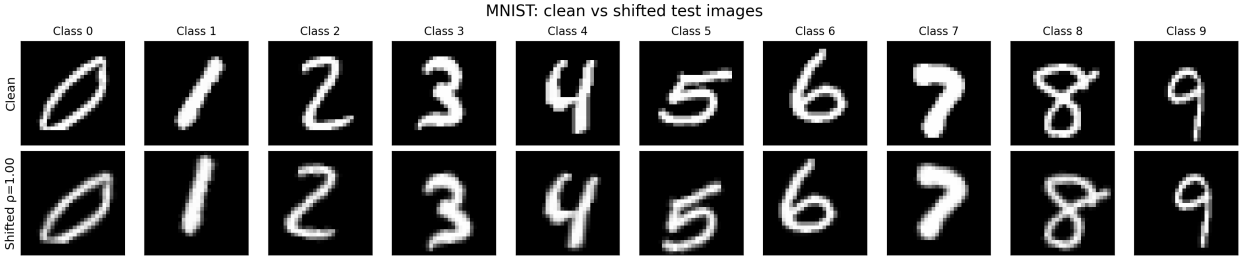


(a) Covariance Stability vs Shift Strength for STL-10 Dataset in CSP1



(b) Covariance Stability vs Shift Strength for STL-10 Dataset in CSP2

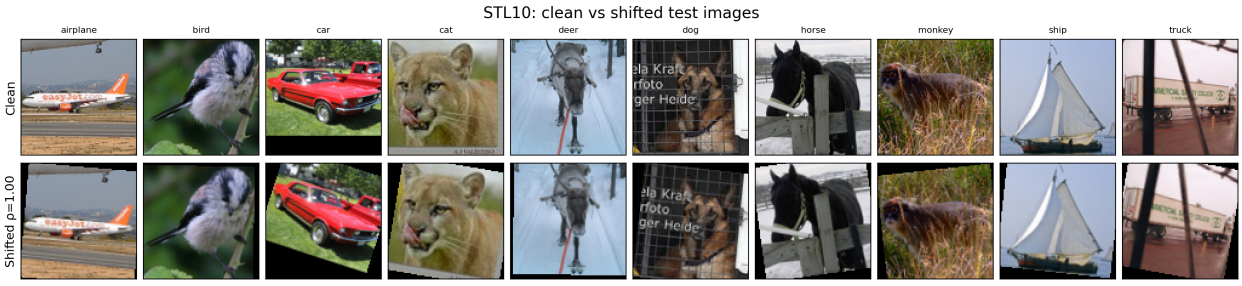
Figure 10: Latent covariance stability, measured by the Frobenius distance $\Delta_\mu(\rho) = \|\widehat{C}_{\text{train}} - \widehat{C}_{\text{test}}(\rho)\|_F$ vs. shift strength ρ for all VIB variants on STL-10. Top: CSP1. Bottom: CSP2.



(a) Test images for MNIST dataset under CSP1

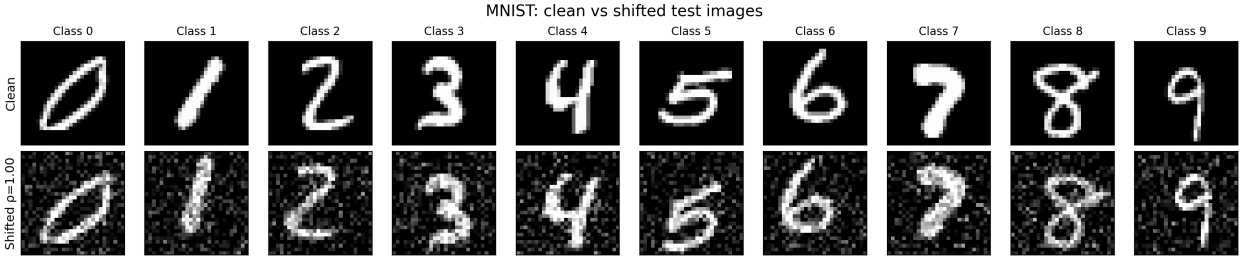


(b) Test images for Cifar-10 dataset under CSP1

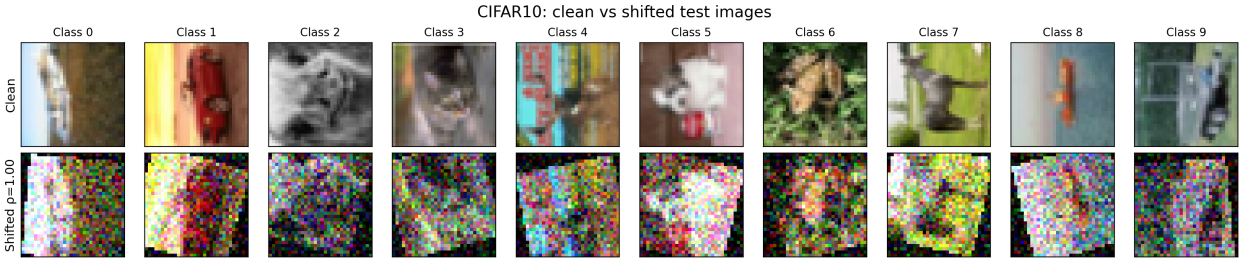


(c) Test images for STL-10 dataset under CSP1

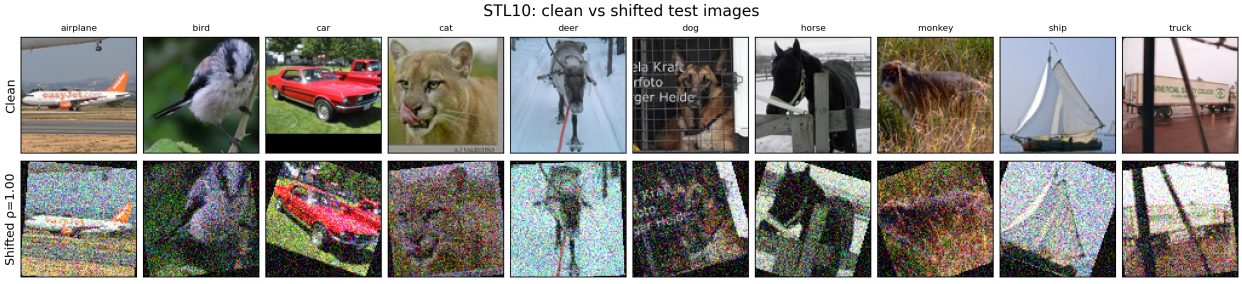
Figure 11: Sample testing images under CSP1.



(a) Test images for MNIST dataset under CSP2



(b) Test images for of Cifar-10 dataset under CSP2



(c) Test images for of STL-10 dataset under CSP2

Figure 12: Sample testing images under CSP2.