VARIATIONAL MANIFOLD ALIGNMENT FOR NOISY MULTILABEL PREDICTION CORRECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Label noise is ubiquitous in machine learning applications and usually leads to a geometric displacement in a trained classifier's latent space. We propose LSNPC (Latent-Shift Noisy Prediction Correction), a post-hoc correction method that refines the multilabel predictions of a classifier trained on noisy labels, without retraining it. LSNPC is a variational autoencoder that aligns clean and noisy label distributions on a shared latent manifold by decoding both through the same decoder; this alignment is the primary corrective mechanism, which we design so that decoding a noisy prediction recovers most of the lost signal. A heavy-tailed Student-t transition models the noise-induced displacement, and a Gaussian shift from the noisy toward the clean representation then closes the residual displacement, yielding a smaller secondary refinement. We prove that the learned correction posterior approximates the true posterior over clean representations, that the Student-t shift yields a Kullback-Leibler divergence bound linear in the label discrepancy (a Gaussian alternative adds a quadratic term), and that the label-covariance reconstruction error is controlled by the latent displacement and label sparsity. Experiments on four synthetic-noise benchmarks, with real-world evaluation in the appendix, show that LSNPC improves prediction performance in most settings, with the largest gains observed under severe corruption.

1 INTRODUCTION

Multilabel classification, where each instance carries multiple labels, poses distinct challenges for learning under noise. Label correlations let a single error cascade across related categories, sparsity drowns the signal, and imbalance makes rare-label errors disproportionately costly (Bogatinovski et al., 2022; Liu et al., 2017). Consider a deployed multilabel image classifier trained on a large dataset with imperfect human annotations, where labels may be either incorrect or missing. Retraining on corrected labels is expensive. Underlying these errors is a geometric mechanism, since label noise is the visible consequence of a latent displacement, not an arbitrary corruption. We therefore study correcting a trained classifier's predictions without retraining it by reversing this displacement.

Learning with noisy labels has advanced considerably in multiclass classification, spanning noise cleansing, sample selection, robust losses, and transition-matrix correction (Song et al., 2022; Han et al., 2018; Wei et al., 2020; Ma et al., 2020; Zhang & Sabuncu, 2018; Wang et al., 2019; Xia et al., 2021; Patrini et al., 2017; Li et al., 2021a; Yang et al., 2022). Most relevant to our setting is the post-processing paradigm of Noise Prediction Correction (NPC) (Bae et al., 2022), which calibrates classifier predictions without retraining. The multilabel setting remains underexplored, since these multiclass methods do not extend directly under label correlations, sparsity, and co-occurrence constraints (Hou et al., 2025).

We study noisy multilabel prediction through the lens of learned latent representations. Our central hypothesis is that label noise induces geometric structure in a learned latent space. Clean labels are generated from a latent variable $\mathbf{z}$, while noisy predictions arise from a perturbed version $\hat{\mathbf{z}}$ connected to $\mathbf{z}$ through a Student-t transition; the label noise is thus a consequence of the latent shift, since the noisy prediction $\hat{\mathbf{y}}$ is generated from the displaced latent $\hat{\mathbf{z}}$ through the same decoder that produces $\mathbf{y}$ from $\mathbf{z}$. This geometric perspective suggests a natural learning objective, namely training a Variational Autoencoder (VAE) (Kingma & Welling, 2014) with a *shared* decoder that requires clean and noisy representations to inhabit the same manifold. Decoding a noisy prediction

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.

through this shared decoder then corrects it; within the aligned manifold, a further displacement from $\hat{\mathbf{z}}$ toward $\mathbf{z}$ refines the correction.

This design yields three insights. First, the shared decoder enforces manifold alignment. Because both clean and noisy labels are decoded through the same network, proximity in latent space must correspond to label similarity for both clean and noisy inputs. Second, the heavy-tailed Student-t shift underlying this correction is not merely a robustness heuristic, since we prove that its KL cost grows at most linearly in the label discrepancy ($\mathcal{O}(\Delta)$, with no quadratic term), whereas a Gaussian alternative carries an additional quadratic term ($\mathcal{O}(m\Delta+m\Delta^2)$). Third, the latent geometry directly controls resilience to sparsity, because the expected covariance error is bounded by the expected latent displacement and by the concentration of the decoder output.

Several in-training methods address the multilabel case. Li et al. (2022) estimate label transition matrices using label correlations. MCM (Gonzalez & Ibrahim, 2025) (Multi-label Confusion Mixture) models inter-label dependencies probabilistically. HLC (Xia et al., 2023) (Holistic Label Correction) exploits instance-label and label-label dependencies during training. UNM (Chen et al., 2024) (a universal approach to noisy multi-label learning) uses label-wise embeddings with cyclical fitting. BalanceMix (Song et al., 2024) refines labels at label-wise granularity. The generative approaches closest to ours are LRAD (Chen et al., 2023) and CAD (Hou et al., 2025). LRAD reformulates noisy-label learning as a generative process that gradually refines an initial random guess with a diffusion model to produce labels, with inference proceeding by maximum-likelihood estimation. CAD similarly treats image features as conditions and the multilabel matrix as the diffusion target, constraining the reverse process with label co-occurrence. Neither LRAD nor CAD is a direct competitor to LSNPC; both are in-training methods that learn the labeling function from noisy data, whereas LSNPC is a post-hoc corrector trained on the base classifier's predictions, serving any pre-trained classifier without retraining it.

Our contributions are threefold. First, we introduce LSNPC (Latent-Shift Noisy Prediction Correction), a VAE that corrects noisy multilabel predictions by learning a shared clean/noisy manifold, with a residual latent displacement as a refinement; one trained model per dataset serves any pre-trained classifier on it. Second, we prove that the unsupervised objective approximates the Bayesian posterior over clean latents, that the KL between clean and noisy posteriors is $\mathcal{O}(\Delta)$ in the label discrepancy under a Student-t shift (no quadratic term, versus a mixed linear–quadratic rate for a Gaussian), and that the covariance reconstruction error is controlled by the latent displacement and label sparsity. Third, extensive experiments show that LSNPC improves base predictions in most settings, most markedly under severe corruption.

2 Method

We consider multilabel classification with k labels. Let $\mathbf{x} \in \mathcal{X} \subseteq \mathbb{R}^d$ be a data point and $\mathbf{y} \in \mathcal{Y} = \{0,1\}^k$ the true label vector. The dataset is $\mathcal{D} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^n$, with noisy observations $\tilde{\mathcal{D}} = \{(\mathbf{x}_i, \tilde{\mathbf{y}}_i)\}_{i=1}^n$. Throughout, D_{KL} denotes the Kullback-Leibler divergence. We study post-hoc correction of a pre-trained classifier $h : \mathcal{X} \mapsto \mathcal{Y}$ trained on $\tilde{\mathcal{D}}$. Its prediction follows $p(\hat{\mathbf{y}} \mid \mathbf{x}, h) \equiv p(\hat{\mathbf{y}} \mid \mathbf{x}, \tilde{\mathcal{D}}) =: p_h(\hat{\mathbf{y}} \mid \mathbf{x})$. Relative to the raw noisy labels $\tilde{\mathbf{y}}$, $p_h(\hat{\mathbf{y}} \mid \mathbf{x})$ provides a refined proxy amenable to Bayesian modeling. Our goal is to learn a correction function $\mathcal{C}(\cdot)$ that maps inputs to corrected labels, using learned representations of the noise process.

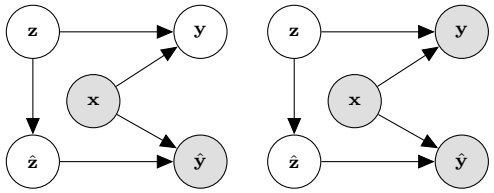
2.1 Modeling Noise-Induced Displacement

Our central hypothesis is that label noise manifests as a geometric displacement in a learned latent space. A shared decoder exploits this structure by enforcing manifold alignment. Each data point $\mathbf{x}$ has an associated clean latent representation $\mathbf{z}$ that generates correct labels $\mathbf{y}$. When labels are noisy, the classifier produces predictions $\hat{\mathbf{y}}$ from a perturbed representation $\hat{\mathbf{z}}$ that is connected to $\mathbf{z}$ through a Student-t transition. The displacement between $\hat{\mathbf{z}}$ and $\mathbf{z}$ encodes the effect of noise, with small displacements yielding reliable predictions and large displacements signaling corruption.

We learn this representation through a VAE. An encoder maps $(\mathbf{x}, \hat{\mathbf{y}})$ to the noisy latent $\hat{\mathbf{z}}$, and a *shared* decoder reconstructs labels from either $\mathbf{z}$ or $\hat{\mathbf{z}}$ using the same network. This shared decoder is the key design choice, enforcing manifold alignment by requiring clean and noisy representations

to live on the same latent manifold. Decoding a prediction through the shared decoder already corrects it, because clean and noisy representations are organized around the same label semantics; moving from $\hat{\mathbf{z}}$ toward $\mathbf{z}$ in the latent space provides a further refinement at inference time.

We model the transition $\mathbf{z} \to \hat{\mathbf{z}}$ as a Student-t distribution rather than a Gaussian. This choice is motivated by representation robustness, in that heavy tails keep large perturbations (severe noise) plausible under the model while small perturbations (mild noise) remain well represented. We later show that this choice yields a KL bound whose form is linear in the label discrepancy, whereas a Gaussian alternative carries an additional quadratic term (Section 3).



(a) Unsupervised (b) Semi-supervised

Figure 1: The process of noisy label generation. Gray indicates observed variables. Left: unsupervised ($\mathbf{y}$ latent). Right: semi-supervised ($\mathbf{y}$ observed for a clean subset).

2.2 Generative Process

The generative process (Figure 1) is formalized as

$$\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \quad \hat{\mathbf{z}} \mid \mathbf{z} \sim \mathrm{T}_{\nu_0}(\mu_\psi(\mathbf{z}), \mathbf{I}), \quad \mathbf{y} \mid \mathbf{z}, \mathbf{x} \sim \mathrm{Bern}(g_\phi(\mathbf{x}, \mathbf{z})), \quad \hat{\mathbf{y}} \mid \hat{\mathbf{z}}, \mathbf{x} \sim \mathrm{Bern}(g_\phi(\mathbf{x}, \hat{\mathbf{z}})).$$

The clean latent $\mathbf{z}$ has a standard Normal prior $\mathcal{N}(\mathbf{0}, \mathbf{I})$, and the noisy latent $\hat{\mathbf{z}}$ follows a Student-t distribution with ν_0 degrees of freedom around the shifted mean $\mu_\psi(\mathbf{z})$. Labels are Bernoulli per component, and the shared decoder g_ϕ, with parameters $\Phi_d = \{\psi, \phi\}$, produces both $\mathbf{y}$ (from $\mathbf{z}$) and $\hat{\mathbf{y}}$ (from $\hat{\mathbf{z}}$). The heavy-tailed property aids robustness (Gao et al., 2021; Li et al., 2021b), and reparameterization is well-established (Takahashi et al., 2018; Abiri & Ohlsson, 2019).

2.3 Unsupervised Learning

Since $\mathbf{y}$ is unobserved, we introduce a variational posterior $q(\mathbf{z}, \hat{\mathbf{z}} \mid \mathbf{x}, \hat{\mathbf{y}})$. Following the VAE framework (Kingma & Welling, 2014), we minimize the KL divergence to the true posterior

$$\begin{aligned} D_{\mathrm{KL}}(q(\mathbf{z}, \hat{\mathbf{z}} \mid \mathbf{x}, \hat{\mathbf{y}}) \,||\, p(\mathbf{z}, \hat{\mathbf{z}} \mid \mathbf{x}, \hat{\mathbf{y}})) &= \mathrm{CONST} - \mathbb{E}_{\hat{\mathbf{z}} \sim q(\hat{\mathbf{z}} \mid \mathbf{x}, \hat{\mathbf{y}})}[\log p(\hat{\mathbf{y}} \mid \mathbf{x}, \hat{\mathbf{z}})] \\ &\quad - \mathbb{E}_{\mathbf{z}, \hat{\mathbf{z}} \sim q}[\log p(\hat{\mathbf{z}} \mid \mathbf{z}) + \log p(\mathbf{z})] + \mathbb{E}_{\hat{\mathbf{z}} \sim q}[\log q(\hat{\mathbf{z}} \mid \mathbf{x}, \hat{\mathbf{y}})] + \mathbb{E}_{\mathbf{z}, \hat{\mathbf{z}} \sim q}[\log q(\mathbf{z} \mid \hat{\mathbf{z}})], \end{aligned} \tag{1}$$

where $q(\mathbf{z}, \hat{\mathbf{z}} \mid \mathbf{x}, \hat{\mathbf{y}}) = q(\mathbf{z} \mid \hat{\mathbf{z}}) q(\hat{\mathbf{z}} \mid \mathbf{x}, \hat{\mathbf{y}})$. Up to a constant, this is the negative evidence lower bound (ELBO), combining the reconstruction term $-\mathbb{E}[\log p(\hat{\mathbf{y}} \mid \mathbf{x}, \hat{\mathbf{z}})]$ with two KL-regularization terms, $D_{\mathrm{KL}}(q(\hat{\mathbf{z}} \mid \mathbf{x}, \hat{\mathbf{y}}) \,||\, p(\hat{\mathbf{z}}))$ and $\mathbb{E}[D_{\mathrm{KL}}(q(\mathbf{z} \mid \hat{\mathbf{z}}) \,||\, p(\mathbf{z}))]$, where $p(\mathbf{z}) = \mathcal{N}(\mathbf{0}, \mathbf{I})$ is the standard-normal prior and $p(\hat{\mathbf{z}})$ is the corresponding generative marginal over noisy latents. Since the decoder outputs a per-label Bernoulli distribution, the reconstruction term is the negative binary cross-entropy (BCE) summed over the k labels, $\log p(\hat{\mathbf{y}} \mid \mathbf{x}, \hat{\mathbf{z}}) = -\mathrm{BCE}(g_\phi(\mathbf{x}, \hat{\mathbf{z}}), \hat{\mathbf{y}})$; it is thus the reconstruction objective of the ELBO and the natural likelihood for the conditionally independent per-component labels. Following the β-VAE scheme (Higgins et al., 2017), we weight the two KL terms by a small β to preserve reconstruction fidelity; the value follows from the mean-based reconstruction scaling described in Section E. We parameterize the encoder distributions as

$$q(\mathbf{z} \mid \hat{\mathbf{z}}) = \mathcal{N}(\mu_\kappa(\hat{\mathbf{z}}), \mathrm{diag}(\sigma_\kappa^2(\hat{\mathbf{z}}))) \quad q(\hat{\mathbf{z}} \mid \mathbf{x}, \hat{\mathbf{y}}) = \mathrm{T}_\nu(\mu_\theta(\mathbf{x}, \hat{\mathbf{y}}), \mathrm{diag}(\sigma_\theta^2(\mathbf{x}, \hat{\mathbf{y}}))). \tag{2}$$

The encoder parameters are $\Phi_e = \{\kappa, \theta\}$. The correction-path mean μ_κ is distinct from the generative shift network μ_ψ; μ_ψ parameterizes the forward transition $\hat{\mathbf{z}} \mid \mathbf{z}$ used to *generate* noisy latents, whereas μ_κ parameterizes the inference-time mapping $\hat{\mathbf{z}} \to \mathbf{z}$ used for correction. The regularity assumptions in Section 3 concern the latter.

The two variational posteriors divide the correction between them. The noisy posterior $q(\hat{\mathbf{z}} \mid \mathbf{x}, \hat{\mathbf{y}})$ is Student-t; its heavy tails let the encoder place mass over the region of the shared manifold from which the decoder reconstructs the true labels, even when $\hat{\mathbf{y}}$ is severely corrupted, rather than collapsing onto the noisy mode. Decoding $\hat{\mathbf{z}}$ through the shared decoder is therefore already corrective, and the direct-$\hat{\mathbf{z}}$ ablation (Section H.4) shows this pathway carries the large majority of the gain. The clean posterior $q(\mathbf{z} \mid \hat{\mathbf{z}})$ is Gaussian and concentrated near $\mu_\kappa(\hat{\mathbf{z}})$, a local recentering of $\hat{\mathbf{z}}$ toward the clean mode that refines the correction rather than producing it. This division is the inference counterpart of the generative story; the Student-t transition $\mathbf{z} \to \hat{\mathbf{z}}$ models the heavy-tailed displacement that noise injects, and the heavy-tailed posterior is what lets inference absorb that displacement, leaving only a residual mean shift for the Gaussian to close. The unsupervised loss over $\tilde{\mathcal{D}}$ is

$$\mathcal{U}(\tilde{\mathcal{D}}) \coloneqq \textstyle\sum_{(\mathbf{x}, _) \in \tilde{\mathcal{D}},\ \hat{\mathbf{y}} \sim h(\mathbf{x})} D_{\mathrm{KL}}(q(\mathbf{z}, \hat{\mathbf{z}} \mid \mathbf{x}, \hat{\mathbf{y}}) \,||\, p(\mathbf{z}, \hat{\mathbf{z}} \mid \mathbf{x}, \hat{\mathbf{y}})). \tag{3}$$

2.4 Supervised Learning

In the supervised setting, where clean labels $\mathbf{y}$ are observed on a subset, we consider the variational posterior $q(\mathbf{z}, \hat{\mathbf{z}} \mid \mathbf{x}, \mathbf{y}, \hat{\mathbf{y}})$ over the two latents. Applying the same KL derivation as in the unsupervised case yields the supervised objective

$$\begin{aligned} &D_{\mathrm{KL}}(q(\mathbf{z}, \hat{\mathbf{z}} \mid \mathbf{x}, \mathbf{y}, \hat{\mathbf{y}}) \,||\, p(\mathbf{z}, \hat{\mathbf{z}} \mid \mathbf{x}, \mathbf{y}, \hat{\mathbf{y}})) \\ &= -\mathbb{E}_{\mathbf{z}, \hat{\mathbf{z}} \sim q(\mathbf{z}, \hat{\mathbf{z}} \mid \mathbf{x}, \mathbf{y}, \hat{\mathbf{y}})}[\log p(\hat{\mathbf{y}} \mid \mathbf{x}, \hat{\mathbf{z}}) + \log p(\mathbf{y} \mid \mathbf{x}, \mathbf{z})] - \mathbb{E}_{\mathbf{z}, \hat{\mathbf{z}} \sim q(\mathbf{z}, \hat{\mathbf{z}} \mid \mathbf{x}, \mathbf{y}, \hat{\mathbf{y}})}[\log p(\hat{\mathbf{z}} \mid \mathbf{z}) + \log p(\mathbf{z})] \\ &\quad + \mathbb{E}_{\hat{\mathbf{z}} \sim q(\hat{\mathbf{z}} \mid \mathbf{x}, \hat{\mathbf{y}})}[\log q(\hat{\mathbf{z}} \mid \mathbf{x}, \hat{\mathbf{y}})] + \mathbb{E}_{\mathbf{z} \sim q(\mathbf{z} \mid \mathbf{x}, \mathbf{y}, \hat{\mathbf{z}})}[\log q(\mathbf{z} \mid \mathbf{x}, \mathbf{y}, \hat{\mathbf{z}})] + \mathrm{CONST.}, \end{aligned} \tag{4}$$

where the variational distribution factorizes as $q(\mathbf{z}, \hat{\mathbf{z}} \mid \mathbf{x}, \mathbf{y}, \hat{\mathbf{y}}) = q(\mathbf{z} \mid \mathbf{x}, \mathbf{y}, \hat{\mathbf{z}}) q(\hat{\mathbf{z}} \mid \mathbf{x}, \hat{\mathbf{y}})$. We specify the clean-latent proposal as a mixture

$$q(\mathbf{z} \mid \mathbf{x}, \mathbf{y}, \hat{\mathbf{z}}) = \eta \, \mathcal{N}(\mu_\theta(\mathbf{x}, \mathbf{y}), \mathrm{diag}(\sigma_\theta^2(\mathbf{x}, \mathbf{y}))) + (1 - \eta) \, q(\mathbf{z} \mid \hat{\mathbf{z}}), \tag{5}$$

while the noisy proposal $q(\hat{\mathbf{z}} \mid \mathbf{x}, \hat{\mathbf{y}}) = \mathrm{T}_\nu(\mu_\theta(\mathbf{x}, \hat{\mathbf{y}}), \mathrm{diag}(\sigma_\theta^2(\mathbf{x}, \hat{\mathbf{y}})))$ is identical to the unsupervised case. The mixture in Equation (5) preserves the correction-time distribution $q(\mathbf{z} \mid \hat{\mathbf{z}})$ while letting the clean subset refine the encoder, a supervision signal the purely unsupervised configuration lacks. The mixing coefficient η trades off these objectives. Even at $\eta = 0$, when $\mathbf{z}$ is always drawn from $q(\mathbf{z} \mid \hat{\mathbf{z}})$, the reconstruction term $\log p(\mathbf{y} \mid \mathbf{x}, \mathbf{z})$ in the supervised loss evaluates the decoder at the corrected latent $\mathbf{z}$ against the true labels, and this reconstruction gradient carries the clean-label signal to the correction path, as the sensitivity study in Section E.3 shows. That gradient propagates along the signal path $\mathrm{BCE}(g_\phi(\mathbf{x}, \mathbf{z}), \mathbf{y}) \to \mathbf{z} \to \mu_\kappa \to \hat{\mathbf{z}} \to \mu_\theta$, hence clean labels inform the correction mapping without a dedicated clean encoder branch. The loss on the clean subset $\mathcal{D}$ is

$$\mathcal{L}(\mathcal{D}) \coloneqq \textstyle\sum_{(\mathbf{x}, \mathbf{y}) \in \mathcal{D}, \, \hat{\mathbf{y}} \sim h(\mathbf{x})} D_{\mathrm{KL}}(q(\mathbf{z}, \hat{\mathbf{z}} \mid \mathbf{x}, \mathbf{y}, \hat{\mathbf{y}}) \,||\, p(\mathbf{z}, \hat{\mathbf{z}} \mid \mathbf{x}, \mathbf{y}, \hat{\mathbf{y}})). \tag{6}$$

2.5 Semi-Supervised Learning

Following the paradigm of Bayesian semi-supervised learning (Kingma et al., 2014), the semi-supervised objective combines the supervised and unsupervised losses as $\min_\Phi \mathcal{L}(\mathcal{D}) + \mathcal{U}(\tilde{\mathcal{D}})$, with $\Phi = \Phi_d \cup \Phi_e$. We adopt the two-phase schedule of Section E.2, in which each epoch alternates an unsupervised pass over the noisy dataset with a supervised pass over the clean subset.

2.6 Correction through the Shared Manifold

Once the representation is learned, correction reduces to a single forward pass, in which we sample $\hat{\mathbf{z}} \sim q(\hat{\mathbf{z}} \mid \mathbf{x}, \hat{\mathbf{y}})$, the clean latent $\mathbf{z} \sim q(\mathbf{z} \mid \hat{\mathbf{z}})$, and decode. The correction function $\mathcal{C}(\mathbf{x}, \hat{\mathbf{y}})$ of Section 2 is realized as the expectation

$$\mathcal{C}(\mathbf{x}, \hat{\mathbf{y}}) \coloneqq \mathbf{y}^* = \mathbb{E}_{\mathbf{y} \sim p(\mathbf{y} \mid \mathbf{x}, \mathbf{z}), \, \mathbf{z} \sim q(\mathbf{z} \mid \hat{\mathbf{z}}), \, \hat{\mathbf{z}} \sim q(\hat{\mathbf{z}} \mid \mathbf{x}, \hat{\mathbf{y}}), \, \hat{\mathbf{y}} \sim p_h(\hat{\mathbf{y}} \mid \mathbf{x})}[\mathbf{y}],$$

approximated by Monte Carlo sampling. No iterative optimization is required, since the geometry learned during training directly encodes the clean/noisy correspondence.

3 Geometric Properties of the Learned Representation

We establish three properties: (i) the unsupervised objective provably approximates the Bayesian posterior over $\mathbf{z}$, (ii) the representation is robust to label perturbations, with the Student-t shift yielding a KL bound of $\mathcal{O}(\Delta)$ in the label discrepancy (no quadratic term), whereas a Gaussian alternative exhibits a mixed linear–quadratic rate, and (iii) the representation's geometry directly controls resilience to label sparsity (see Sections A, B and D for proofs).

3.1 Posterior Approximation Guarantee

Theorem 3.1. *Minimizing the unsupervised loss in Equation (3) minimizes an upper bound on the expected KL divergence between the proposed posterior $q(\mathbf{z} \mid \hat{\mathbf{z}})$ and the true conditional posterior $p(\mathbf{z} \mid \hat{\mathbf{z}}, \mathbf{x}, \hat{\mathbf{y}})$. Formally, for a given $\mathbf{x}$,*

$$\mathbb{E}_{\hat{\mathbf{z}} \sim q(\hat{\mathbf{z}} \mid \mathbf{x}, \hat{\mathbf{y}})} \left[D_{\mathrm{KL}}(q(\mathbf{z} \mid \hat{\mathbf{z}}) \,||\, p(\mathbf{z} \mid \hat{\mathbf{z}}, \mathbf{x}, \hat{\mathbf{y}})) \right] \leq D_{\mathrm{KL}}(q(\mathbf{z}, \hat{\mathbf{z}} \mid \mathbf{x}, \hat{\mathbf{y}}) \,||\, p(\mathbf{z}, \hat{\mathbf{z}} \mid \mathbf{x}, \hat{\mathbf{y}})). \tag{7}$$

This follows directly from the chain rule for KL divergence, which gives $D_{\mathrm{KL}}(q(\mathbf{z}, \hat{\mathbf{z}} \mid \mathbf{x}, \hat{\mathbf{y}}) \,||\, p(\mathbf{z}, \hat{\mathbf{z}} \mid \mathbf{x}, \hat{\mathbf{y}})) = D_{\mathrm{KL}}(q(\hat{\mathbf{z}} \mid \mathbf{x}, \hat{\mathbf{y}}) \,||\, p(\hat{\mathbf{z}} \mid \mathbf{x}, \hat{\mathbf{y}})) + \mathbb{E}_{\hat{\mathbf{z}} \sim q}[D_{\mathrm{KL}}(q(\mathbf{z} \mid \hat{\mathbf{z}}) \,||\, p(\mathbf{z} \mid \hat{\mathbf{z}}, \mathbf{x}, \hat{\mathbf{y}}))]$. Since both terms on the right are non-negative, the inequality holds without further assumptions. This justifies using $q(\mathbf{z} \mid \hat{\mathbf{z}})$ in the correction phase; the learned representation provably approximates the Bayesian posterior over $\mathbf{z}$ induced by the noisy latent $\hat{\mathbf{z}}$ (formal proof in Section A).

3.2 Representation Robustness to Label Noise

Let $\Delta(\cdot, \cdot)$ be a symmetric metric on $\mathcal{Y}$ with $\Delta(\mathbf{y}_0, \mathbf{y}_1) = 0$ iff $\mathbf{y}_0 = \mathbf{y}_1$. We introduce mild continuity assumptions on the encoder.

Assumption 3.2 (Encoder regularity)**.** Let $\mathcal{S}_\theta$ denote the parameter space of the encoder over which the conditions below are required to hold. The encoder satisfies three regularity conditions:

(i) The encoder log-variance is Lipschitz in the label. For all $\mathbf{x} \in \mathcal{X}$, $\theta \in \mathcal{S}_\theta$, and $\mathbf{y}_0, \mathbf{y}_1 \in \mathcal{Y}$ with $\Delta(\mathbf{y}_0, \mathbf{y}_1) > 0$, there exists $M > 0$ such that $|\log \sigma^2_\theta(\mathbf{x}, \mathbf{y}_0)_j - \log \sigma^2_\theta(\mathbf{x}, \mathbf{y}_1)_j| \leq M\Delta(\mathbf{y}_0, \mathbf{y}_1)$ for all j.

(ii) The encoder mean is Lipschitz continuous w.r.t. the label metric, $\|\mu_\theta(\mathbf{x}, \mathbf{y}_0) - \mu_\theta(\mathbf{x}, \mathbf{y}_1)\|_2 \leq L_\theta \Delta(\mathbf{y}_0, \mathbf{y}_1)$.

(iii) Encoder variance is bounded away from zero uniformly over inputs, labels, and encoder parameters, with $\lambda \leq \inf_{\mathbf{x} \in \mathcal{X},\, \mathbf{y} \in \mathcal{Y},\, \theta \in \mathcal{S}_\theta} \min_j \sigma^2_\theta(\mathbf{x}, \mathbf{y})_j$.

These conditions ensure that small label perturbations induce only bounded encoder changes; the next result transfers this continuity to the KL divergence between the clean and noisy posteriors.

Theorem 3.3. *Let m be the latent dimension. Under Theorem 3.2, for true label $\mathbf{y}_0$ and noisy label $\mathbf{y}_1$ with $\nu > 2$ and $\Delta(\mathbf{y}_0, \mathbf{y}_1) \in [0, 1]$,*

$$D_{\mathrm{KL}}(q(\mathbf{z} \mid \mathbf{x}, \hat{\mathbf{y}} = \mathbf{y}_1) \,||\, q(\mathbf{z} \mid \mathbf{x}, \hat{\mathbf{y}} = \mathbf{y}_0)) = \mathcal{O}\big(\Delta(\mathbf{y}_0, \mathbf{y}_1)\big), \tag{8}$$

where $q(\mathbf{z} \mid \mathbf{x}, \hat{\mathbf{y}} = \mathbf{y}_i) := \int q(\mathbf{z} \mid \hat{\mathbf{z}})\, q(\hat{\mathbf{z}} \mid \mathbf{x}, \hat{\mathbf{y}} = \mathbf{y}_i)\, d\hat{\mathbf{z}}$ is the clean-latent marginal induced by the two-stage encoder, and $\mathcal{O}(\cdot)$ is taken with respect to Δ at fixed model parameters. The bound is affine in Δ, with no quadratic term, up to a label-independent proof-slack intercept; explicit constants and the full proof appear in Section B.

Under a Student-t shift, the KL between clean and noisy posteriors therefore grows at most linearly, $\mathcal{O}(\Delta)$, whereas a Gaussian proposal gives the rate $\mathcal{O}(m\Delta + m\Delta^2)$ proved in Section C. The two bounds differ in the quadratic term; it is absent for the Student-t (Theorem B.5) and present for the Gaussian with coefficient $mL^2_\theta/(2\lambda)$ on top of a linear coefficient $m\tilde{M}$, where $\tilde{M} := Me^M$ and M is the log-Lipschitz constant of Assumption 3.2 (i). Within $\Delta \in [0, 1]$ the quadratic Gaussian term can exceed its linear term when $L^2_\theta \Delta/(2\lambda) > \tilde{M}$, hence the comparison concerns functional form rather than pointwise ordering. The representation-choice ablation in Section H.3 is consistent with this distinction; the gap widens as noise severity increases.

3.3 Sparsity and Imbalance

A defining challenge of multilabel learning is extreme label sparsity. We now show that the bound on the corrected covariance error tightens as the representation concentrates mass on fewer labels, formalizing a structural interplay between sparsity and latent geometry.

Definition 3.4 (Weighted Effective Dimension)**.** For a nonzero probability vector $\boldsymbol{p} \in [0, 1]^k$, the weighted effective dimension is defined as $d_w(\boldsymbol{p}) = \|\boldsymbol{p}\|^2_1 / \|\boldsymbol{p}\|^2_2$; it satisfies $1 \leq d_w(\boldsymbol{p}) \leq k$, approaching 1 for vectors concentrated on few coordinates and k for uniform ones.

Definition 3.5 (Effective Dimension)**.** For $\epsilon > 0$ and a probability vector $\boldsymbol{p}$ with sorted coordinates $p_{(1)} \geq p_{(2)} \geq \cdots \geq p_{(k)}$, the effective dimension is $d_\epsilon(\boldsymbol{p}) = \min\{\ell : \sum_{i=1}^{\ell} p_{(i)} \geq 1 - \epsilon\}$.

Assumption 3.6 (Decoder regularity and sparsity structure)**.** The decoder and the correction path satisfy four structural conditions:

(i) The decoder g_ϕ and the correction mean μ_κ are Lipschitz: $\|g_\phi(\mathbf{x}, \mathbf{z}_1) - g_\phi(\mathbf{x}, \mathbf{z}_2)\|_2 \leq L_\phi \|\mathbf{z}_1 - \mathbf{z}_2\|_2$, $\|\mu_\kappa(\mathbf{z}_1) - \mu_\kappa(\mathbf{z}_2)\|_2 \leq L_\kappa \|\mathbf{z}_1 - \mathbf{z}_2\|_2$.

(ii) The decoded probabilities are bounded, $\pi_{\min} \leq [g_\phi(\mathbf{x}, \mathbf{z})]_i \leq \pi_{\max}$ for all $\mathbf{x}, \mathbf{z}, i$; the proof only uses the closed-interval fact $|1 - q_i - p_i| \leq 1$ for $q_i, p_i \in [0, 1]$.

(iii) The correction path is bounded, $\|\mu_\kappa(\boldsymbol{\mu}_{\hat{\mathbf{z}}}) - \boldsymbol{\mu}_{\hat{\mathbf{z}}}\|_2 \leq R$, where $\boldsymbol{\mu}_{\hat{\mathbf{z}}} = \mu_\theta(\mathbf{x}, \hat{\mathbf{y}})$ is the noisy latent mean.

(iv) The decoded outputs have non-degenerate weighted effective dimension (Theorem 3.4), $d_w(g_\phi(\mathbf{x}, \mathbf{z})) \geq d_{\min}$ and $d_w(g_\phi(\mathbf{x}, \hat{\mathbf{z}})) \geq d_{\min}$ for some $d_{\min} \geq 1$.

The direct-$\hat{\mathbf{z}}$ decoding ablation (Section H.4) shows that the explicit correction shift is small in practice (0.2–1.3 CF1 points across configurations), consistent with a modest R in Assumption 3.6 (iii). A well-aligned shared manifold already places $\hat{\mathbf{z}}$ near the clean region, leaving only a small displacement for the correction path to close.

Theorem 3.7 (Two-Sided Covariance Characterization)**.** *Under Theorem 3.6, let* $\mathbf{y}_r \sim \mathrm{Bern}(g_\phi(\mathbf{x}, \mathbf{z}))$ *and* $\hat{\mathbf{y}}_r \sim \mathrm{Bern}(g_\phi(\mathbf{x}, \hat{\mathbf{z}}))$*. Then the following two-sided bound holds*

$$\epsilon_{\text{shift}} - \epsilon_{\text{recon}} - C_{\text{latent}} \leq \mathbb{E}\left[\|\boldsymbol{\Sigma}_{\mathbf{y}} - \boldsymbol{\Sigma}_{\mathbf{y}_r}\|_F\right] \leq \epsilon_{\text{shift}} + \epsilon_{\text{recon}} + C_{\text{latent}}, \tag{9}$$

where ϵ_{shift} *is the intrinsic noise,* ϵ_{recon} *the reconstruction error, and* $C_{\text{latent}} = \left(2\rho/\sqrt{d_{\min}} + 1\right) L_\phi \left(\sigma_{\mathbf{z}} + (L_\kappa + 1)\sigma_{\hat{\mathbf{z}}} + R\right)$ *with* $\rho = d_\epsilon \pi_{\max} + \epsilon$*,* $\sigma_{\mathbf{z}} = \sqrt{\mathbb{E}_{\hat{\mathbf{z}}}[\mathrm{tr}(\boldsymbol{\Sigma}_{\mathbf{z}})]}$*, and* $\sigma_{\hat{\mathbf{z}}} = \sqrt{\frac{\nu}{\nu-2}\,\mathrm{tr}(\boldsymbol{\Sigma}_{\hat{\mathbf{z}}})}$*, where* $\boldsymbol{\Sigma}_{\mathbf{z}} = \mathrm{diag}(\sigma_\kappa^2(\hat{\mathbf{z}}))$ *and* $\boldsymbol{\Sigma}_{\hat{\mathbf{z}}} = \mathrm{diag}(\sigma_\theta^2(\mathbf{x}, \hat{\mathbf{y}}))$ *are the conditional covariance matrices of the variational posteriors* $q(\mathbf{z} \mid \hat{\mathbf{z}})$ *and* $q(\hat{\mathbf{z}} \mid \mathbf{x}, \hat{\mathbf{y}})$ *defined in Section 2.*

The proof appears in Section D. The quantity $\rho = d_\epsilon \pi_{\max} + \epsilon$ shrinks with the effective dimension d_ϵ, tightening C_{latent} since sparser label vectors offer fewer opportunities for noise to create spurious co-occurrences. Since $C_{\text{latent}} = \mathcal{O}(\rho/\sqrt{d_{\min}})$, the bound degrades only sublinearly in the diversity constraint $d_{\min}$. Theorem 3.7 thus formalizes a structural link between sparsity and representation quality, the latent component of the covariance-error bound tightening as the per-instance label distribution concentrates. Empirically this holds on VOC12, where the covariance-error reduction is largest in the sparsest bin. On Tomato, however, the corrected covariance error increases in every reported bin (Section H.2); the advantage is therefore dataset-conditional. We caution that the net tightening depends on the relative rates at which d_ϵ and d_w concentrate, which we do not quantify; the empirical evidence is consistent with the tighter bound on sparse instances. The two-sided characterization pins the limit of post-hoc correction. Whenever the intrinsic noise ϵ_{shift} exceeds the recovery ability $\epsilon_{\text{recon}} + C_{\text{latent}}$, the corrected covariance error is bounded below by a strictly positive quantity; hence LSNPC cannot reduce it below the floor set by the label noise that the base classifier inherits. This lower bound is conditional and is tightest where the intrinsic noise dominates. Sparsity also narrows the band, since C_{latent} tightens as labels concentrate.

The same bound has a natural interpretation for label-wise imbalance; the weighted effective dimension $d_w(\boldsymbol{p})$ measures concentration, with $d_w \to 1$ for a vector dominated by few coordinates (whether from instance-wise sparsity or a skewed label-frequency distribution) and $d_w \to k$ for a diffuse one. The diversity constraint $d_w(g_\phi(\mathbf{x}, \cdot)) \geq d_{\min}$ (Assumption 3.6 (iv)) then prevents the latent component of the bound from collapsing. This is an interpretation, and a dataset-level claim about imbalance or minority-class recovery would require assumptions we do not make.

Together, Theorem 3.1, Theorem 3.3, and Theorem 3.7 characterize the learned representation, and are supported by the ablations in Section H.3 and Section H.2.

4 Experiments

As our experiments study post-hoc correction of noisy models, we include only baseline methods we could reliably reproduce under our protocol. We evaluate the learned representations along three axes: whether the shared latent manifold improves predictions across diverse noise regimes; whether access to a small clean set further refines the representation; and whether ablations support the geometric claims about sparsity and the Student-t shift.

4.1 Experimental Design

Datasets and noise. We evaluate on VOC07 and VOC12 (Everingham, 2007), COCO (Lin et al., 2014), and Tomato (Gehlot et al., 2023) in Section H.1. Noise is generated via class-dependent

transition matrices (Xia et al., 2023; Patrini et al., 2017) with symmetric (Sym) and pairflip (Pair) noise (Van Rooyen et al., 2015; Han et al., 2018) at 30%, 40%, and 50% rates. For VOC12 and COCO, 30% noise results are omitted from the main figures; all four datasets include results at 40% and 50% noise. The 50% noise level serves as a stress test, an unrealistic setting for deployed systems, yet one that starkly reveals the asymptotic properties of the representation.

Baselines and metrics. We evaluate LSNPC as a post-hoc correction on top of a set of base predictors spanning three encoder architectures: ResNet-50 (He et al., 2016) (subscript r), Vision Transformer fine-tuned at 224×224 resolution (ViT-224, Dosovitskiy et al., 2021; subscript v), and Swin-T (Wu et al., 2020) (subscript s). On VOC07 we instantiate eight base classifiers and six on Tomato. On ResNet-50 these are the standard multilabel classifier MLP_r, the graph-based ADDGCN_r (Ye et al., 2020), and the noise-aware HLC_r (Xia et al., 2023) and UNM_r (Chen et al., 2024). On Swin-T we use MLP again (MLP_s, isolating the backbone effect), together with the noise-aware MCM_s (Gonzalez & Ibrahim, 2025), the asymmetric-loss classifier ASL_s (Ridnik et al., 2021), and LRAD_s (Chen et al., 2023), a diffusion-based label generation baseline (marked † in tables). On VOC12 and COCO, we evaluate five base predictors: the standard MLP_v, the asymmetric-loss ASL_v, and the noise-aware HLC_r, MCM_v, and UNM_v. Each base predictor appears both alone and with LSNPC applied (+LSNPC). The multilabel extension of NPC degenerates to a trivial, noise-invariant fixed point in every configuration (Section F). We report per-class F1 (CF1), overall F1 (OF1) (Guo et al., 2025; Xia et al., 2023; Song et al., 2024), and mean Average Precision (mAP). All LSNPC models use a β-VAE (Higgins et al., 2017) with $\beta = 0.001$, latent dimension $m = 32$, and degrees of freedom $\nu = \nu_0 = 2.001$. The mixing coefficient is set to $\eta = 0.1$ (see Section E.3 for justification). We average results over five independent runs per configuration and report mean $\pm$ standard error. Full implementation details appear in Section E. We provide the code for our method and experiments as supplemental material. A more capable image encoder for LSNPC generally yields larger improvements over the base model; for a fair comparison, we nevertheless match the LSNPC encoder to that of each base model.

4.2 Experimental Results

Table 1: Unsupervised correction on VOC07. Blue: LSNPC improves over baseline. Values are reported as baseline / +LSNPC, with the latter given as Mean $\pm$ SE. †: generative baseline.

Noise Rate	Baseline / +LSNPC	Symmetric			Pairflip		
		CF1	OF1	mAP	CF1	OF1	mAP
30%	MLP_r / +LSNPC	77.2 / **79.7**±0.5	79.0 / **80.9**±0.4	87.8 / 87.8±0.1	76.6 / **77.4**±0.5	79.0 / **79.6**±0.5	82.7 / **83.0**±0.2
	ADDGCN_r / +LSNPC	74.6 / **77.7**±0.6	75.9 / **78.6**±1.0	87.3 / **87.5**±0.1	72.6 / **75.5**±0.6	75.4 / **77.6**±0.6	82.8 / **83.0**±0.2
	HLC_r / +LSNPC	79.1 / **80.7**±0.4	80.9 / **82.0**±0.4	88.0 / 87.9±0.1	78.0 / **79.4**±0.3	80.6 / **81.0**±0.3	86.0 / 85.7±0.0
	UNM_r / +LSNPC	74.2 / **76.6**±0.5	76.0 / **78.2**±0.4	85.5 / **85.9**±0.3	61.0 / **65.9**±1.0	62.6 / **66.9**±0.9	82.5 / **83.6**±0.3
	MLP_s / +LSNPC	82.9 / **83.3**±0.6	81.3 / **82.6**±1.0	90.6 / **90.7**±0.3	82.9 / **83.4**±0.2	84.3 / **84.8**±0.2	87.7 / 87.5±0.2
	MCM_s / +LSNPC	79.8 / **82.6**±0.5	81.3 / **83.3**±0.5	90.1 / **90.4**±0.3	75.8 / **79.7**±0.4	78.5 / **82.3**±0.2	84.9 / **87.0**±0.3
	ASL_s / +LSNPC	83.1 / 82.1±0.9	84.6 / 83.4±0.5	91.1 / 89.4±0.3	81.8 / 81.0±0.5	83.6 / 82.5±0.3	87.2 / 85.7±0.4
	$\mathrm{LRAD}_s^{\dagger}$ / +LSNPC	83.7 / **84.6**±0.1	82.6 / **83.8**±0.4	90.5 / **90.9**±0.1	85.0 / **87.4**±0.3	85.5 / **85.6**±0.3	88.8 / **89.0**±0.2
40%	MLP_r / +LSNPC	70.6 / **75.6**±0.4	72.5 / **76.6**±0.7	86.9 / 86.7±0.1	65.3 / 65.1±1.0	69.6 / 69.2±0.8	73.0 / **74.7**±0.8
	ADDGCN_r / +LSNPC	64.2 / **72.1**±1.0	67.8 / **74.4**±1.1	86.6 / 86.2±0.1	56.7 / **61.2**±0.6	63.3 / **66.9**±0.3	69.3 / **71.6**±0.2
	HLC_r / +LSNPC	75.4 / **78.0**±0.6	77.6 / **79.5**±0.7	87.3 / 86.9±0.1	67.2 / **68.1**±0.7	71.9 / **73.0**±0.6	81.0 / **81.2**±0.3
	UNM_r / +LSNPC	65.6 / **70.9**±0.9	69.2 / **73.4**±0.8	84.1 / **84.5**±0.1	35.2 / **36.3**±0.6	38.3 / **40.4**±1.0	68.9 / **70.8**±0.4
	MLP_s / +LSNPC	73.7 / **79.1**±0.4	78.0 / **81.3**±0.4	90.4 / 90.2±0.2	71.2 / **71.4**±1.2	73.5 / **74.3**±1.4	79.3 / **80.1**±0.7
	MCM_s / +LSNPC	73.0 / **79.2**±0.9	72.2 / **78.1**±0.5	88.8 / **90.0**±0.4	63.8 / **69.7**±0.4	67.5 / **72.5**±0.7	73.6 / **80.2**±0.1
	ASL_s / +LSNPC	77.5 / **78.4**±0.9	79.9 / **80.3**±0.5	90.1 / 88.2±0.4	68.2 / 67.9±1.8	72.4 / 69.5±1.4	77.2 / **77.6**±0.7
	$\mathrm{LRAD}_s^{\dagger}$ / +LSNPC	82.2 / **83.4**±0.6	82.9 / **84.1**±0.5	89.6 / **90.3**±0.2	70.2 / **71.3**±0.6	69.2 / **71.0**±1.0	82.6 / **82.9**±0.3
50%	MLP_r / +LSNPC	58.7 / **67.4**±1.1	61.5 / **69.8**±1.3	84.7 / 84.6±0.2	46.2 / **48.3**±2.0	47.4 / **49.4**±2.6	53.6 / **55.9**±0.3
	ADDGCN_r / +LSNPC	44.7 / **55.7**±1.1	48.8 / **59.8**±0.8	85.1 / 83.8±0.2	44.3 / **48.7**±1.2	47.2 / **50.0**±1.7	53.1 / **55.5**±0.7
	HLC_r / +LSNPC	71.6 / **74.5**±0.7	73.5 / **75.5**±0.8	85.4 / 85.1±0.1	43.1 / **44.8**±1.1	43.4 / **45.8**±1.0	54.1 / **58.3**±0.7
	UNM_r / +LSNPC	46.8 / **54.5**±1.4	50.9 / **58.2**±2.2	80.5 / **81.1**±0.3	19.8 / **21.6**±0.2	17.4 / **18.5**±0.3	53.6 / **54.9**±0.5
	MLP_s / +LSNPC	49.2 / **59.1**±1.6	57.5 / **66.6**±1.3	88.3 / 88.0±0.2	37.8 / **48.4**±0.4	38.5 / **50.2**±1.0	56.9 / **59.4**±0.4
	MCM_s / +LSNPC	58.8 / **66.7**±0.6	58.5 / **67.5**±0.9	87.5 / 87.4±0.4	47.0 / **49.3**±0.4	54.0 / **56.0**±0.4	54.4 / **57.6**±0.6
	ASL_s / +LSNPC	60.8 / **62.9**±2.4	63.3 / **68.0**±2.3	87.6 / 84.8±0.9	45.3 / **49.4**±0.2	47.7 / **53.6**±0.6	56.5 / **57.9**±0.2
	$\mathrm{LRAD}_s^{\dagger}$ / +LSNPC	68.7 / **74.7**±1.2	71.8 / **76.7**±1.0	86.4 / **88.0**±0.3	46.9 / **50.6**±0.9	43.8 / **48.2**±1.9	62.4 / **65.1**±0.8

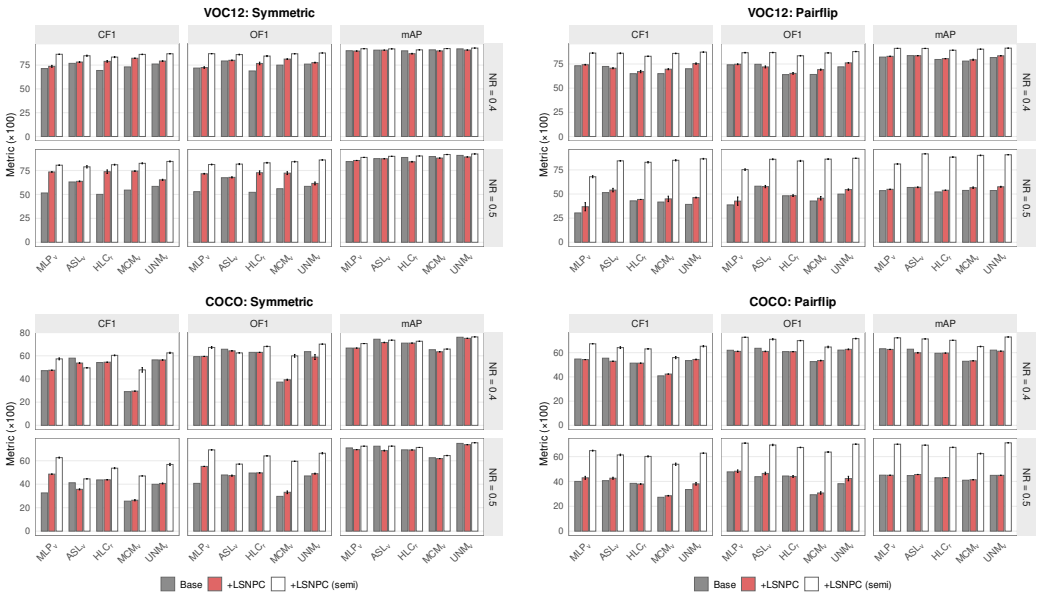


Figure 2: Semi-supervised results on VOC12 and COCO at 40% and 50% noise rates (NR). A small clean set (5.6% of the training partition on VOC12 and 12.2% on COCO; see Table 2) further refines the learned representation, improving CF1, OF1, and mAP.

Unsupervised representation learning. Table 1 reports results on VOC07 across all eight baselines. LSNPC improves CF1 and OF1 in most configurations. The gains are largest at 50% noise, a deliberately extreme regime, where the base classifiers lose substantial predictive performance and LSNPC recovers much of the lost signal (e.g., $+8.7$ CF1 for MLP$_r$ at Sym-50%, and $+10.6$ CF1 for MLP$_s$ at Pair-50%). The gains are smaller at 30% noise, and in a few settings LSNPC produces a modest regression, concentrated on the asymmetric-loss classifier (ASL$_s$ loses about one CF1 point at Sym-30% and Pair-30%, with isolated sub-point regressions on MLP$_r$ at Pair-40%). This pattern is consistent with the theoretical analysis; the Student-t shift's bound has no quadratic term in Δ, hence its advantage over a Gaussian alternative grows with the label discrepancy (Theorem 3.3), and under mild noise the correction gain is accordingly modest.

Semi-supervised representation learning. Figure 2 shows that a small clean subset substantially boosts the correction. On COCO at 50% noise, semi-supervised LSNPC lifts corrected predictions to levels comparable to the base classifiers at 40% noise, effectively recovering a full noise-rate step, and the gains are large and consistent on VOC12 and across both noise types. The improvement spans all three metrics; the semi-supervised arm performs best in nearly every configuration. The mAP trajectory is the most nuanced of the three; the unsupervised LSNPC is neutral to slightly negative, particularly under symmetric noise where the base ranking is largely intact, and it is the clean subset that lifts mAP decisively above the base. This mirrors the threshold-versus-ranking analysis of Section 4.2 and indicates that the clean subset restores within-class ranking rather than only sharpening thresholds. ASL, whose base predictions are strongest, is the notable exception; the unsupervised correction degrades them on COCO, most visibly under symmetric noise, while producing only small mixed changes on VOC12. At COCO Sym-40% even the semi-supervised variant stays below the base on all three metrics; at 50% noise it recovers above the base on CF1 and OF1 while only matching it on mAP. This is consistent with an asymmetric-loss classifier whose scores are already well calibrated and offer little room for post-hoc correction. Where the base is weakest under severe noise, the semi-supervised gains are largest, with the corrected arms recovering most of the lost performance. This is a distinctive capability, since none of the baselines is designed for semi-supervised training (their objectives are defined solely on the noisy dataset).

Analysis of mAP. A recurring pattern in Table 1 is that LSNPC improves CF1 and OF1 in nearly all settings, while mAP exhibits mixed behavior, improving under pairflip corruption and degrading

under moderate symmetric noise when the base classifier already performs well. The discrepancy traces to an objective mismatch. The ELBO optimizes pointwise reconstruction fidelity (covariance error, per Theorem 3.7) rather than within-class ranking, and the Bernoulli decoder, combined with weak KL regularization, drives output probabilities toward 0 or 1, a phenomenon known as confidence sharpening (Guo et al., 2017). Crossing the 0.5 threshold directly improves CF1 and OF1, while the concentrated outputs can scramble within-class ordering and degrade mAP, a ranking metric (Section G gives a numerical example). Across the VOC07 configurations, symmetric noise leaves Δ_{mAP} near zero at 30–40%, while pairflip noise improves it by a few points on average (largest on MCM_s Pair-40% at $+6.6$, where the base ranking was already poor). LSNPC is therefore most impactful under high noise or weak base classifiers; under mild noise, strong noise-aware classifiers may already be the better choice for ranking-sensitive applications. The threshold-versus-ranking gap is analogous to the pointwise-versus-listwise distinction in learning-to-rank, suggesting that ranking-aware objectives could reconcile the two metric families.

4.3 Ablations

The remaining ablations are summarized here; details are deferred to the appendix. Two ablations verify the geometric claims of Section 3. The sparsity ablation (Section H.2) measures the Label Covariance Norm (LCN) per label-density bin. On VOC12 (MLP_r, Sym-50%), LSNPC reduces LCN in every density bin, with the largest reductions on the sparsest instances (Table 5), supporting the two-sided covariance characterization of Theorem 3.7. The representation-choice ablation (Section H.3 and Table 6) confirms Theorem 3.3; the Student-t representation outperforms a Gaussian alternative, with the gap widening under severe noise. On VOC12 (ADDGCN, Sym-50%), the Gaussian variant yields CF1 49.4 versus 70.7 for the Student-t.

A further ablation validates the shared-decoder design (Section H.5). Replacing the shared decoder with one per branch collapses the correction; on VOC07 (MLP_r, 50% noise), the separate-decoder variant attains CF1 of 11.2 (mean over the repetitions), against 48.3 (pairflip) and 67.4 (symmetric) CF1 for the shared decoder (the same repetition runs as Table 1). The collapse is structural; the unsupervised objective supervises only the noisy branch $\hat{\mathbf{y}}$ from $\hat{\mathbf{z}}$; a decoder that never receives that signal cannot learn to correct. The shared decoder is precisely what transfers the noisy-supervision signal to the clean branch, enabling correction. Together with the direct-$\hat{\mathbf{z}}$ ablation below, this collapse identifies the shared decoder as the source of the correction and the latent shift as a residual refinement.

Finally, one ablation bypasses the $\mathbf{z} \sim q(\mathbf{z} \mid \hat{\mathbf{z}})$ step, testing whether the shift is essential rather than incidental. Across the VOC12 MLP_r, UNM_r, and HLC_r configurations (Section H.4), such direct-$\hat{\mathbf{z}}$ decoding captures most of the correction, with the full LSNPC adding only a small gain. In the configurations with substantial correction, direct-$\hat{\mathbf{z}}$ decoding alone recovers roughly 84–98% of the total CF1 gain, the explicit $\hat{\mathbf{z}} \to \mathbf{z}$ shift contributing the remainder; decoding through the shared decoder is thus the dominant corrective pathway, while the shift's modest secondary gain is largest under severe symmetric corruption. On the real-world DeepFashion benchmark (Section H.6), LSNPC improves mAP for all six baselines, confirming that the correction transfers beyond synthetic noise.

5 Conclusion

We introduced LSNPC, a VAE that corrects noisy multilabel predictions by learning a shared manifold for clean and noisy labels. The shared decoder is the primary corrective mechanism, forcing clean and noisy representations onto the same manifold; hence decoding a noisy prediction already recovers most of the lost signal, while a Gaussian shift closes the residual displacement as a smaller secondary refinement. Extensive experiments show gains in most settings, largest under severe noise, with further improvement from a semi-supervised variant exploiting a small clean set. The shared-manifold principle suggests the approach may generalize beyond multilabel classification to other structured prediction tasks, with ranking-aware objectives and instance-conditional noise models as the most promising future directions. For future work, one may use the posterior over clean latents as an uncertainty estimate for the corrected labels, flagging the predictions least amenable to correction and prioritizing them for annotation.

AI USE STATEMENT

In this work, we used generative AI tools for editing the paper to improve readability. We have not used generative AI tools for developing theoretical models or conceptual frameworks, formulating mathematical claims, providing critical ingredients for proving mathematical claims, assisting in the writing of proofs, proposing or refining hypotheses, designing or providing feedback on research methodology or experiments, implementing methods, assisting with translation, cleaning and reformatting datasets, supporting qualitative and thematic data analysis, or interpreting results. Generating synthetic data sets is not applicable to this work. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

ACKNOWLEDGMENTS

REFERENCES

Najmeh Abiri and Mattias Ohlsson. Variational auto-encoders with student's t-prior. In *ESANN 2019 - Proceedings*, 2019. 27th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, ESANN 2019 ; Conference date: 24-04-2019 Through 26-04-2019.

HeeSun Bae, Seungjae Shin, Byeonghu Na, JoonHo Jang, Kyungwoo Song, and Il-Chul Moon. From noisy prediction to true label: Noisy prediction calibration via generative model. In *International Conference on Machine Learning*, 5 2022.

Jasmin Bogatinovski, Ljupčo Todorovski, Sašo Džeroski, and Dragi Kocev. Comprehensive comparative study of multi-label classification methods. *Expert Systems with Applications*, 203:117215, 2022. ISSN 0957-4174. doi: https://doi.org/10.1016/j.eswa.2022.117215.

Jia-Yao Chen, Shao-Yuan Li, Sheng-Jun Huang, Songcan Chen, Lei Wang, and Ming-Kun Xie. Unm: A universal approach for noisy multi-label learning. *IEEE Transactions on Knowledge and Data Engineering*, 2024.

Jian Chen, Ruiyi Zhang, Tong Yu, Rohan Sharma, Zhiqiang Xu, Tong Sun, and Changyou Chen. Label-retrieval-augmented diffusion models for learning from noisy labels. In *Advances in Neural Information Processing Systems*, volume 36, pp. 66499–66517, 2023. arXiv:2305.19518.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

Mark Everingham. The pascal visual object classes challenge,(voc2007) results. *http://pascallin. ecs. soton. ac. uk/challenges/VOC/voc2007/index. html.*, 2007.

Wei Gao, Teng Zhang, Bin-Bin Yang, and Zhi-Hua Zhou. On the noise estimation statistics. *Artificial Intelligence*, 293:103451, 2021.

Mamta Gehlot, Rakesh Kumar Saxena, and Geeta Chhabra Gandhi. "tomato-village": a dataset for end-to-end tomato disease detection in a real-world environment. *Multimedia Systems*, 29(6): 3305–3328, 2023.

Diego Linares Gonzalez and Shahana Ibrahim. Multi-label recognition under noisy supervision: A confusion mixture modeling approach. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2025.

Benjamin Graham, Alaaeldin El-Nouby, Hugo Touvron, Pierre Stock, Armand Joulin, Hervé Jégou, and Matthijs Douze. Levit: a vision transformer in convnet's clothing for faster inference. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 12259–12269, 2021.

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

Jiawei Guo, Yang Xiao, Weipeng Huang, and Guangyuan Piao. Enhancing text-based hierarchical multilabel classification for mobile applications via contrastive learning. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pp. 4476–4486, 2025.

Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. *Advances in neural information processing systems*, 31, 2018.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *International conference on learning representations*, 2017.

Senyu Hou, Yuru Ren, Gaoxia Jiang, and Wenjian Wang. Noisy multi-label learning through co-occurrence-aware diffusion. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Yulong Huang, Yonggang Zhang, and Jonathon A. Chambers. A novel kullback–leibler divergence minimization-based adaptive student's t-filter. *IEEE Transactions on Signal Processing*, 67(20): 5417–5432, 2019. doi: 10.1109/TSP.2019.2939079.

Diederik P. Kingma and Max Welling. Auto-Encoding Variational Bayes. In *2nd International Conference on Learning Representations, ICLR*, 2014.

Durk P Kingma, Shakir Mohamed, Danilo Jimenez Rezende, and Max Welling. Semi-supervised learning with deep generative models. *Advances in neural information processing systems*, 27, 2014.

Shikun Li, Xiaobo Xia, Hansong Zhang, Yibing Zhan, Shiming Ge, and Tongliang Liu. Estimating noise transition matrix with label correlations for noisy multi-label learning. *Advances in Neural Information Processing Systems*, 35:24184–24198, 2022.

Xuefeng Li, Tongliang Liu, Bo Han, Gang Niu, and Masashi Sugiyama. Provably end-to-end label-noise learning without anchor points. In *International conference on machine learning*, pp. 6403–6413. PMLR, 2021a.

Yiming Li, Yang Zhang, Qingtao Tang, Weipeng Huang, Yong Jiang, and Shu-Tao Xia. t-k-means: A robust and stable k-means variant. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3120–3124, 2021b.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pp. 740–755. Springer, 2014.

Weiwei Liu, Ivor W Tsang, and Klaus-Robert Müller. An easy-to-hard learning paradigm for multiple classes and multiple labels. *Journal of Machine Learning Research*, 18(94):1–38, 2017.

Xingjun Ma, Hanxun Huang, Yisen Wang, Simone Romano, Sarah Erfani, and James Bailey. Normalized loss functions for deep learning with noisy labels. In *International conference on machine learning*, pp. 6543–6553. PMLR, 2020.

Roger B Nelsen. *An introduction to copulas*. Springer, 2006.

Giorgio Patrini, Alessandro Rozza, Aditya Krishna Menon, Richard Nock, and Lizhen Qu. Making deep neural networks robust to label noise: A loss correction approach. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1944–1952, 2017.

Tal Ridnik, Emanuel Ben-Baruch, Nadav Zamir, Asaf Noy, Itamar Friedman, Matan Protter, and Lihi Zelnik-Manor. Asymmetric loss for multi-label classification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 82–91, 2021.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pp. 618–626, 2017.

Hwanjun Song, Minseok Kim, Dongmin Park, Yooju Shin, and Jae-Gil Lee. Learning from noisy labels with deep neural networks: A survey. *IEEE transactions on neural networks and learning systems*, 34(11):8135–8153, 2022.

Hwanjun Song, Minseok Kim, and Jae-Gil Lee. Toward robustness in multi-label classification: A data augmentation strategy against imbalance and noise. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 21592–21601, 2024.

Hiroshi Takahashi, Tomoharu Iwata, Yuki Yamanaka, Masanori Yamada, and Satoshi Yagi. Student-t variational autoencoder for robust density estimation. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pp. 2696–2702. International Joint Conferences on Artificial Intelligence Organization, 7 2018. doi: 10.24963/ijcai.2018/374.

MTCAJ Thomas and A Thomas Joy. *Elements of Information Theory*. Wiley-Interscience, 2006.

Brendan Van Rooyen, Aditya Menon, and Robert C Williamson. Learning with symmetric label noise: The importance of being unhinged. *Advances in neural information processing systems*, 28, 2015.

Xi Wang and Junming Yin. Relaxed multivariate bernoulli distribution and its applications to deep generative models. In *Conference on Uncertainty in Artificial Intelligence*, pp. 500–509. PMLR, 2020.

Yisen Wang, Xingjun Ma, Zaiyi Chen, Yuan Luo, Jinfeng Yi, and James Bailey. Symmetric cross entropy for robust learning with noisy labels. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 322–330, 2019.

Hongxin Wei, Lei Feng, Xiangyu Chen, and Bo An. Combating noisy labels by agreement: A joint training method with co-regularization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 13726–13735, 2020.

Bichen Wu, Chenfeng Xu, Xiaoliang Dai, Alvin Wan, Peizhao Zhang, Zhicheng Yan, Masayoshi Tomizuka, Joseph Gonzalez, Kurt Keutzer, and Peter Vajda. Visual transformers: Token-based image representation and processing for computer vision. *arXiv preprint arXiv:2006.03677*, 2020.

Xi-Zhu Wu and Zhi-Hua Zhou. A unified view of multi-label performance measures. In *International Conference on Machine Learning*, pp. 3780–3788. PMLR, 2017.

Xiaobo Xia, Tongliang Liu, Bo Han, Chen Gong, Nannan Wang, Zongyuan Ge, and Yi Chang. Robust early-learning: Hindering the memorization of noisy labels. In *International Conference on Learning Representations*, 2021.

Xiaobo Xia, Jiankang Deng, Wei Bao, Yuxuan Du, Bo Han, Shiguang Shan, and Tongliang Liu. Holistic label correction for noisy multi-label classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1483–1493, 2023.

Shuo Yang, Erkun Yang, Bo Han, Yang Liu, Min Xu, Gang Niu, and Tongliang Liu. Estimating instance-dependent bayes-label transition matrix using a deep neural network. In *International Conference on Machine Learning*, pp. 25302–25312. PMLR, 2022.

Jin Ye, Junjun He, Xiaojiang Peng, Wenhao Wu, and Yu Qiao. Attention-driven dynamic graph convolutional network for multi-label image recognition. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXI 16*, pp. 649–665. Springer, 2020.

Zhilu Zhang and Mert Sabuncu. Generalized cross entropy loss for training deep neural networks with noisy labels. *Advances in neural information processing systems*, 31, 2018.

A Proof of Theorem 3.1

Proof. The chain rule for KL divergence states that for any joint distributions $q(z, w)$ and $p(z, w)$,

$$D_{\mathrm{KL}}(q(z, w) \,||\, p(z, w)) = D_{\mathrm{KL}}(q(w) \,||\, p(w)) + \mathbb{E}_{w\sim q}\left[D_{\mathrm{KL}}(q(z \mid w) \,||\, p(z \mid w))\right].$$

Applying this with $z = \mathbf{z}$, $w = \hat{\mathbf{z}}$ under the variational and true posteriors (with $\mathbf{x}, \hat{\mathbf{y}}$ held fixed) gives

$$\begin{aligned} D_{\mathrm{KL}}(q(\mathbf{z}, \hat{\mathbf{z}} \mid \mathbf{x}, \hat{\mathbf{y}}) \,||\, p(\mathbf{z}, \hat{\mathbf{z}} \mid \mathbf{x}, \hat{\mathbf{y}})) &= D_{\mathrm{KL}}(q(\hat{\mathbf{z}} \mid \mathbf{x}, \hat{\mathbf{y}}) \,||\, p(\hat{\mathbf{z}} \mid \mathbf{x}, \hat{\mathbf{y}})) \\ &\quad + \mathbb{E}_{\hat{\mathbf{z}}\sim q(\hat{\mathbf{z}}|\mathbf{x},\hat{\mathbf{y}})}\left[D_{\mathrm{KL}}(q(\mathbf{z} \mid \hat{\mathbf{z}}) \,||\, p(\mathbf{z} \mid \hat{\mathbf{z}}, \mathbf{x}, \hat{\mathbf{y}}))\right] \\ &\geq \mathbb{E}_{\hat{\mathbf{z}}\sim q(\hat{\mathbf{z}}|\mathbf{x},\hat{\mathbf{y}})}\left[D_{\mathrm{KL}}(q(\mathbf{z} \mid \hat{\mathbf{z}}) \,||\, p(\mathbf{z} \mid \hat{\mathbf{z}}, \mathbf{x}, \hat{\mathbf{y}}))\right], \end{aligned} \tag{10}$$

where the inequality drops the non-negative marginal KL term. The left-hand side is the unsupervised loss per data point (Equation (3)), and the right-hand side measures the approximation quality of $q(\mathbf{z} \mid \hat{\mathbf{z}})$. Minimizing $\mathcal{U}$ therefore directly drives down the expected posterior approximation error. Applying the same chain-rule argument to the supervised objective (conditioning additionally on $\mathbf{y}$) yields an analogous bound, confirming that both objectives push $q(\mathbf{z} \mid \hat{\mathbf{z}})$ toward $p(\mathbf{z} \mid \hat{\mathbf{z}}, \mathbf{x}, \hat{\mathbf{y}})$. □

B Proof of Theorem 3.3

Proof. The divergence we must bound compares the clean-latent posteriors $q(\mathbf{z} \mid \mathbf{x}, \hat{\mathbf{y}} = \mathbf{y}_1)$ and $q(\mathbf{z} \mid \mathbf{x}, \hat{\mathbf{y}} = \mathbf{y}_0)$. These arise from a two-stage random-variable flow. First, the encoder maps the label to a noisy latent,

$$\hat{\mathbf{z}} \mid (\mathbf{x}, \hat{\mathbf{y}}) \sim q(\hat{\mathbf{z}} \mid \mathbf{x}, \hat{\mathbf{y}}) = \mathrm{T}_\nu\big(\mu_\theta(\mathbf{x}, \hat{\mathbf{y}}), \mathrm{diag}(\sigma^2_\theta(\mathbf{x}, \hat{\mathbf{y}}))\big), \tag{11}$$

and second, the correction channel draws a clean latent from the noisy one,

$$\mathbf{z} \mid \hat{\mathbf{z}} \sim q(\mathbf{z} \mid \hat{\mathbf{z}}) = \mathcal{N}\big(\mu_\kappa(\hat{\mathbf{z}}), \mathrm{diag}(\sigma^2_\kappa(\hat{\mathbf{z}}))\big), \tag{12}$$

which is independent of $\hat{\mathbf{y}}$ given $\hat{\mathbf{z}}$. The two stages therefore form the Markov chain

$$\hat{\mathbf{y}} \to \hat{\mathbf{z}} \to \mathbf{z}, \tag{13}$$

with $\hat{\mathbf{y}}$ the input label, $\hat{\mathbf{z}}$ the intermediate noisy latent, and $\mathbf{z}$ the leaf clean latent.

The data processing inequality (Thomas & Joy, 2006) states that for any Markov chain $U \to V \to W$, the divergence between the leaf distributions induced by two instantiations u_0, u_1 of the root never exceeds the divergence between the induced intermediate distributions

$$D_{\mathrm{KL}}\big(p(W \mid U = u_1) \,||\, p(W \mid U = u_0)\big) \leq D_{\mathrm{KL}}\big(p(V \mid U = u_1) \,||\, p(V \mid U = u_0)\big), \tag{14}$$

because the channel $V \to W$ is the same stochastic map for both instantiations, and a fixed stochastic map can only shrink KL divergence. Applying this to Equation (13) with $U = \hat{\mathbf{y}}$, $V = \hat{\mathbf{z}}$, $W = \mathbf{z}$, and $u_i = \mathbf{y}_i$, and writing $\pi_i(\hat{\mathbf{z}}) \coloneqq q(\hat{\mathbf{z}} \mid \mathbf{x}, \hat{\mathbf{y}} = \mathbf{y}_i)$ for the induced noisy-latent marginals, yields the key reduction

$$D_{\mathrm{KL}}(q(\mathbf{z} \mid \mathbf{x}, \hat{\mathbf{y}} = \mathbf{y}_1) \,||\, q(\mathbf{z} \mid \mathbf{x}, \hat{\mathbf{y}} = \mathbf{y}_0)) \leq D_{\mathrm{KL}}(\pi_1 \,||\, \pi_0). \tag{15}$$

The right-hand side of Equation (15) compares two Student-t distributions $q(\hat{\mathbf{z}} \mid \mathbf{x}, \hat{\mathbf{y}} = \mathbf{y}_i) = \mathrm{T}_\nu(\boldsymbol{\mu}_{\mathbf{y}_i}, \boldsymbol{\Sigma}_{\mathbf{y}_i})$ with shared ν, hence by Theorem B.3 and Theorem B.4 it is bounded by the affine-in-Δ expression $C_1 + C_2\,\Delta(\mathbf{y}_0, \mathbf{y}_1)$, which contains no quadratic term; the intercept C_1 is label-independent proof slack (Theorem B.5). Combined with Equation (15), the clean-latent divergence is therefore $\mathcal{O}(\Delta(\mathbf{y}_0, \mathbf{y}_1))$ (at most linear in Δ, with no Δ^2 term), as stated in Theorem 3.3. □

B.1 KL-Divergence Between Student-t Distributions

We use the KL-divergence bound between two multivariate Student-t distributions from Huang et al. (2019), restated as Theorem B.1.

Theorem B.1. *Given two multivariate Student-t distributions* $p_1(\mathbf{x}) = \mathrm{T}_{\nu_1}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ *and* $p_2(\mathbf{x}) = \mathrm{T}_{\nu_2}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$ *with* $\nu_1 > 2$ *and* $\tilde{\boldsymbol{\Sigma}}_1 = \frac{\nu_1}{\nu_1 - 2}\boldsymbol{\Sigma}_1$. *The KL divergence between them is*

$$\begin{aligned} D_{\mathrm{KL}}(p_1(\mathbf{x}) \,||\, p_2(\mathbf{x})) \leq \frac{1}{2}\log\frac{\det(\boldsymbol{\Sigma}_2)}{\det(\boldsymbol{\Sigma}_1)} + \frac{m}{2}\log\frac{\nu_2}{\nu_1} + \log\frac{\Gamma(\nu_2/2)}{\Gamma(\nu_1/2)} \\ + \log\frac{\Gamma((\nu_1+m)/2)}{\Gamma((\nu_2+m)/2)} \\ - \frac{\nu_1+m}{2}\left\{\Psi\left((\nu_1+m)/2\right) - \Psi(\nu_1/2)\right\} \\ + \frac{\nu_2+m}{2}\log\left\{1 + \frac{1}{\nu_2}\operatorname{tr}(\boldsymbol{\Sigma}_2^{-1}\tilde{\boldsymbol{\Sigma}}_1) \right. \\ \left. + \frac{1}{\nu_2}\operatorname{tr}\left(\boldsymbol{\Sigma}_2^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top\right)\right\}. \end{aligned} \tag{16}$$

The coefficients $\frac{m}{2}$ on $\log(\nu_2/\nu_1)$ and 1 on the two gamma log-ratios reflect the $\nu^{m/2}$, $\Gamma(\nu/2)$, and $\Gamma((\nu+m)/2)$ factors in the multivariate t density normalizer; they reduce to the familiar $\frac{1}{2}$, $\frac{1}{2}$, and $\frac{1}{2}$ in the univariate case $m = 1$ and vanish when $\nu_1 = \nu_2$ (Theorem B.3).

Remark B.2. The inequality in Theorem B.1 arises from the final step of the standard derivation, where the expectation $\mathbb{E}[\log(1 + D/\nu_2)]$ is bounded above by $\log(1 + \mathbb{E}[D]/\nu_2)$ via Jensen's inequality (since $\log(1+\cdot)$ is concave). Numerical verification with typical parameter values (e.g., $\nu = 2.001$, $m = 32$, random covariances) confirms that the exact KL is strictly less than the bound, with the gap growing as the scale of D increases. For the purposes of this paper, the upper bound suffices; the corollary and all subsequent bounds correctly propagate this inequality.

Corollary B.3. *When* $\nu = \nu_1 = \nu_2$, *the KL simplifies to*

$$\begin{aligned} D_{\mathrm{KL}}(p_1(\mathbf{x}) \,||\, p_2(\mathbf{x})) \leq \frac{1}{2}\log\frac{\det(\boldsymbol{\Sigma}_2)}{\det(\boldsymbol{\Sigma}_1)} - \frac{\nu+m}{2}\left\{\Psi\left(\frac{\nu+m}{2}\right) - \Psi\left(\frac{\nu}{2}\right)\right\} \\ + \frac{\nu+m}{2}\log\left\{1 + \frac{1}{\nu}\operatorname{tr}(\boldsymbol{\Sigma}_2^{-1}\tilde{\boldsymbol{\Sigma}}_1) \right. \\ \left. + \frac{1}{\nu}\operatorname{tr}(\boldsymbol{\Sigma}_2^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top)\right\}. \end{aligned} \tag{17}$$

Lemma B.4. *Under Assumptions 3.2 (i) to 3.2 (iii), for any* $\mathbf{y}_0, \mathbf{y}_1$ *with* $\Delta(\mathbf{y}_0, \mathbf{y}_1) \in [0, 1]$, *the bound is*

$$D_{\mathrm{KL}}(q(\hat{\mathbf{z}} \mid \mathbf{x}, \mathbf{y}_1) \,||\, q(\hat{\mathbf{z}} \mid \mathbf{x}, \mathbf{y}_0)) \leq C_1 + C_2\Delta(\mathbf{y}_0, \mathbf{y}_1), \tag{18}$$

where, letting M *be the log-Lipschitz constant in Assumption 3.2 (i),* $\tilde{M} \coloneqq Me^M$ *(such that* $\exp(M\Delta) \leq 1 + \tilde{M}\Delta$ *for* $\Delta \in [0, 1]$*),* $\alpha = \frac{\sqrt{\nu\lambda}}{L_\theta}$, $A = \frac{\nu+m}{2}$, $S = \sqrt{1 + m/(\nu-2)}$, *and* $D = \Psi\left(\frac{\nu+m}{2}\right) - \Psi\left(\frac{\nu}{2}\right)$,

$$C_1 = A\log S^2 + \frac{A\tilde{M}\sqrt{m}\,\alpha}{2(\nu-2)S} - AD, \tag{19}$$

$$C_2 = \frac{m\tilde{M}}{2} + \frac{A\sqrt{m}}{\alpha S}, \tag{20}$$

and $\Psi(\cdot)$ *denotes the digamma function (logarithmic derivative of the gamma function,* $\Psi(x) = \frac{d}{dx}\log\Gamma(x)$*). The rate is therefore* $\mathcal{O}(\Delta(\mathbf{y}_0, \mathbf{y}_1))$ *with no quadratic term; the label-independent intercept* $C_1 \geq 0$ *is proof slack (Theorem B.5).*

Proof of Theorem B.4. For a given $\mathbf{x}$ and label $\mathbf{y}$, denote $\boldsymbol{\mu}_\mathbf{y} \coloneqq \mu_\theta(\mathbf{x}, \mathbf{y})$ and $\boldsymbol{\Sigma}_\mathbf{y} \coloneqq \operatorname{diag}(\sigma_\theta^2(\mathbf{x}, \mathbf{y}))$. We instantiate Theorem B.3 with $p_1 = q(\hat{\mathbf{z}} \mid \mathbf{x}, \mathbf{y}_1)$ and $p_2 = q(\hat{\mathbf{z}} \mid \mathbf{x}, \mathbf{y}_0)$, both sharing the same degrees of freedom ν, and with $\tilde{\boldsymbol{\Sigma}}(\mathbf{y}_1) = \frac{\nu}{\nu-2}\boldsymbol{\Sigma}_{\mathbf{y}_1}$. Substituting into the

corollary gives

$$D_{\mathrm{KL}}(q(\hat{\mathbf{z}} \mid \mathbf{x}, \mathbf{y}_1) \,||\, q(\hat{\mathbf{z}} \mid \mathbf{x}, \mathbf{y}_0)) \leq \underbrace{\frac{1}{2} \log \frac{\det(\boldsymbol{\Sigma}_{\mathbf{y}_0})}{\det(\boldsymbol{\Sigma}_{\mathbf{y}_1})}}_{\text{(A) determinant ratio}} \underbrace{- \frac{\nu+m}{2} \left\{ \Psi\left(\frac{\nu+m}{2}\right) - \Psi\left(\frac{\nu}{2}\right) \right\}}_{\text{(B) digamma constant}} + \underbrace{\frac{\nu+m}{2} \log \left\{ 1 + \frac{1}{\nu} \operatorname{tr}(\boldsymbol{\Sigma}_{\mathbf{y}_0}^{-1} \tilde{\boldsymbol{\Sigma}}(\mathbf{y}_1)) + \frac{1}{\nu} \operatorname{tr}\left(\boldsymbol{\Sigma}_{\mathbf{y}_0}^{-1} (\boldsymbol{\mu}_{\mathbf{y}_1} - \boldsymbol{\mu}_{\mathbf{y}_0})(\boldsymbol{\mu}_{\mathbf{y}_1} - \boldsymbol{\mu}_{\mathbf{y}_0})^\top\right) \right\}}_{\text{(C) trace-and-mean log term}}.$$

We now bound each component in terms of the label discrepancy $\Delta(\mathbf{y}_0, \mathbf{y}_1)$.

Bounding (A). The eigenvalues of $\boldsymbol{\Sigma}_{\mathbf{y}}$ are the per-dimension variances $\sigma_\theta^2(\mathbf{x}, \mathbf{y})_j$, hence $\det(\boldsymbol{\Sigma}_{\mathbf{y}}) = \prod_{j=1}^m \sigma_{\theta,j}^2$. By Assumption 3.2 (i), $|\log \sigma_{\theta,j}^2(\mathbf{x}, \mathbf{y}_0) - \log \sigma_{\theta,j}^2(\mathbf{x}, \mathbf{y}_1)| \leq M\Delta$, hence the ratio satisfies

$$\frac{\sigma_{\theta,j}^2(\mathbf{x}, \mathbf{y}_0)}{\sigma_{\theta,j}^2(\mathbf{x}, \mathbf{y}_1)} \leq \exp(M\Delta) \leq 1 + \tilde{M}\Delta$$

where $\tilde{M} \coloneqq Me^M$ (using $e^x \leq 1 + xe^x$ for $x \geq 0$, with $\Delta \leq 1$); the effective constant follows directly from the assumption's log-Lipschitz constant via $\tilde{M} = Me^M$. The determinant ratio is therefore bounded by

$$\frac{\det(\boldsymbol{\Sigma}_{\mathbf{y}_0})}{\det(\boldsymbol{\Sigma}_{\mathbf{y}_1})} \leq (1 + \tilde{M}\Delta)^m.$$

Using $\log(1 + x) \leq x$ for all $x \geq 0$ yields the bound

$$\frac{1}{2} \log \frac{\det(\boldsymbol{\Sigma}_{\mathbf{y}_0})}{\det(\boldsymbol{\Sigma}_{\mathbf{y}_1})} \leq \frac{m\tilde{M}}{2} \Delta(\mathbf{y}_0, \mathbf{y}_1).$$

Bounding (C). We treat the two trace terms inside the logarithm separately. For the variance-ratio term, substituting $\tilde{\boldsymbol{\Sigma}}(\mathbf{y}_1) = \frac{\nu}{\nu-2} \boldsymbol{\Sigma}_{\mathbf{y}_1}$ and applying the variance ratio bound $\sigma_{\theta,j}^2(\mathbf{x}, \mathbf{y}_1)/\sigma_{\theta,j}^2(\mathbf{x}, \mathbf{y}_0) \leq 1 + \tilde{M}\Delta$ (derived from Assumption 3.2 (i) via $\exp(M\Delta) \leq 1 + \tilde{M}\Delta$ with $\tilde{M} = Me^M$) elementwise gives

$$\frac{1}{\nu} \operatorname{tr}\left(\boldsymbol{\Sigma}_{\mathbf{y}_0}^{-1} \tilde{\boldsymbol{\Sigma}}(\mathbf{y}_1)\right) = \frac{1}{\nu - 2} \sum_{j=1}^m \frac{\sigma_{\theta,j}^2(\mathbf{x}, \mathbf{y}_1)}{\sigma_{\theta,j}^2(\mathbf{x}, \mathbf{y}_0)} \leq \frac{m(1 + \tilde{M}\Delta)}{\nu - 2} = \frac{m}{\nu - 2} + \frac{\tilde{M}m}{\nu - 2} \Delta(\mathbf{y}_0, \mathbf{y}_1).$$

For the mean-shift term, applying the Lipschitz bound on the encoder mean (Assumption 3.2 (ii)) and the rotational invariance of the trace yields

$$\operatorname{tr}\left(\boldsymbol{\Sigma}_{\mathbf{y}_0}^{-1} (\boldsymbol{\mu}_{\mathbf{y}_0} - \boldsymbol{\mu}_{\mathbf{y}_1})(\boldsymbol{\mu}_{\mathbf{y}_0} - \boldsymbol{\mu}_{\mathbf{y}_1})^\top\right) = (\boldsymbol{\mu}_{\mathbf{y}_0} - \boldsymbol{\mu}_{\mathbf{y}_1})^\top \boldsymbol{\Sigma}_{\mathbf{y}_0}^{-1} (\boldsymbol{\mu}_{\mathbf{y}_0} - \boldsymbol{\mu}_{\mathbf{y}_1}) \leq \frac{mL_\theta^2}{\lambda} \Delta(\mathbf{y}_0, \mathbf{y}_1)^2,$$

where we used the minimum-variance bound λ (Assumption 3.2 (iii)) to control the inverse covariance via $\|\boldsymbol{\Sigma}_{\mathbf{y}_0}^{-1}\|_2 \leq 1/\lambda$, and the Lipschitz constant L_θ to bound the Euclidean mean displacement.

Let $T = \frac{1}{\nu} \operatorname{tr}(\boldsymbol{\Sigma}_{\mathbf{y}_0}^{-1} \tilde{\boldsymbol{\Sigma}}(\mathbf{y}_1)) + \frac{1}{\nu} \operatorname{tr}(\boldsymbol{\Sigma}_{\mathbf{y}_0}^{-1} (\boldsymbol{\mu}_{\mathbf{y}_1} - \boldsymbol{\mu}_{\mathbf{y}_0})(\boldsymbol{\mu}_{\mathbf{y}_1} - \boldsymbol{\mu}_{\mathbf{y}_0})^\top)$ denote the sum inside the logarithm. Substituting the bounds above gives

$$T = \underbrace{\frac{m}{\nu - 2}}_{c} + \underbrace{\frac{\tilde{M}m}{\nu - 2} \Delta + \frac{mL_\theta^2}{\nu\lambda} \Delta^2}_{y = a\Delta + b\Delta^2},$$

where $a = \tilde{M}m/(\nu - 2)$ and $b = mL_\theta^2/(\nu\lambda)$. To obtain a *linear* bound in Δ that is as tight as possible, we exploit the exact multiplicative structure of the logarithm. Since $\log(1 + c + y) = \log(1 + c) + \log\big(1 + y/(1 + c)\big)$ (an identity, not an inequality), we apply the standard inequality

$\log(1+u) \le \sqrt{u}$ for $u \ge 0$ only to the Δ-dependent term $u = y/(1+c)$, and then complete the square on $\sqrt{y/(1+c)}$:

$$\begin{aligned}\frac{\nu+m}{2}\log(1+T) &= \frac{\nu+m}{2}\left\{\log(1+c) + \log\left(1+\frac{y}{1+c}\right)\right\} \\ &\le \frac{\nu+m}{2}\left\{\log(1+c) + \sqrt{\frac{y}{1+c}}\right\} \\ &= \frac{\nu+m}{2}\log(1+c) + \frac{\nu+m}{2\sqrt{1+c}}\sqrt{a\Delta+b\Delta^2} \\ &\le \frac{\nu+m}{2}\log(1+c) + \frac{\nu+m}{2\sqrt{1+c}}\left(\sqrt{b}\,\Delta + \frac{a}{2\sqrt{b}}\right),\end{aligned}$$

where the last step uses $\sqrt{a\Delta+b\Delta^2} \le \sqrt{b}\,\Delta + a/(2\sqrt{b})$ from completing the square ($a\Delta + b\Delta^2 \le (\sqrt{b}\,\Delta + a/(2\sqrt{b}))^2$ where the negative offset $-a^2/(4b)$ has been dropped). Substituting $\alpha = \sqrt{\nu\lambda}/L_\theta$ gives $\sqrt{b} = \sqrt{m}/\alpha$ and $a/(2\sqrt{b}) = \tilde{M}\sqrt{m}\,\alpha/(2(\nu-2))$, hence

$$\text{(C)} \le \underbrace{\frac{\nu+m}{2}\log\left(1+\frac{m}{\nu-2}\right) + \frac{(\nu+m)\tilde{M}\sqrt{m}\,\alpha}{4(\nu-2)\sqrt{1+m/(\nu-2)}}}_{\text{absorbed into } C_1} + \underbrace{\frac{(\nu+m)\sqrt{m}}{2\alpha\sqrt{1+m/(\nu-2)}}}_{\text{contributes to } C_2}\Delta(\mathbf{y}_0,\mathbf{y}_1).$$

This "exact-split" approach is strictly tighter than applying $\log(1+x) \le \sqrt{x}$ globally and then using subadditivity of $\sqrt{\cdot}$, because $\log(1+c) < \sqrt{c}$ and $\sqrt{y/(1+c)} < \sqrt{y}$ whenever $c > 0$. The completing-the-square step converts the quadratic $a\Delta + b\Delta^2$ into a linear $\sqrt{b}\,\Delta$ term; without it, the Δ^2 contribution from the mean-shift trace would persist, and the bound would remain quadratic in Δ rather than reducing to a linear form.

Summing the bounds for (A), (B), and (C) and collecting terms in $\Delta(\mathbf{y}_0,\mathbf{y}_1)$ and constants yields the stated C_1, C_2 in Theorem B.4. □

Remark B.5. The bound in Theorem B.4 is non-vacuous at $\Delta = 0$; there $\mathbf{y}_0 = \mathbf{y}_1$, the exact KL vanishes, and the right-hand side reduces to $C_1 \ge 0$. Positivity follows from Jensen's inequality applied to the concave logarithm. The digamma term is the exact expectation

$$AD = \mathbb{E}\left[\tfrac{\nu+m}{2}\log(1+\Lambda/\nu)\right]$$

where $\Lambda = (\hat{\mathbf{z}} - \boldsymbol{\mu}_{\mathbf{y}_1})^\top \boldsymbol{\Sigma}_{\mathbf{y}_1}^{-1}(\hat{\mathbf{z}} - \boldsymbol{\mu}_{\mathbf{y}_1})$ under $\hat{\mathbf{z}} \sim q(\hat{\mathbf{z}} \mid \mathbf{x}, \mathbf{y}_1)$, and Λ has $\mathbb{E}[\Lambda] = \frac{\nu m}{\nu-2}$, hence

$$AD = \mathbb{E}\left[\tfrac{\nu+m}{2}\log(1+\Lambda/\nu)\right] \le \tfrac{\nu+m}{2}\log\left(1+\tfrac{m}{\nu-2}\right) = A\log S^2,$$

which together with the strictly positive term $A\tilde{M}\sqrt{m}\,\alpha/\big(2(\nu-2)S\big)$ yields $C_1 \ge 0$. Note that the sharper intermediate bound $D \le \log(1+m/\nu)$ does *not* hold in general; the comparison is against $\log(1+m/(\nu-2))$.

C Gaussian Proposal Rate

Corollary C.1 (Gaussian Proposal Rate)**.** *Under Theorem 3.2, if the noisy proposal is a multivariate Normal, $q(\hat{\mathbf{z}} \mid \mathbf{x}, \hat{\mathbf{y}}) = \mathcal{N}(\mu_\theta(\mathbf{x}, \hat{\mathbf{y}}), \mathrm{diag}(\sigma_\theta^2(\mathbf{x}, \hat{\mathbf{y}})))$, then for any $\mathbf{y}_0, \mathbf{y}_1$ with $\Delta(\mathbf{y}_0, \mathbf{y}_1) \in [0, 1]$,*

$$D_{\mathrm{KL}}(q(\hat{\mathbf{z}} \mid \mathbf{x}, \hat{\mathbf{y}} = \mathbf{y}_1) \,||\, q(\hat{\mathbf{z}} \mid \mathbf{x}, \hat{\mathbf{y}} = \mathbf{y}_0)) \le \tilde{M} m\,\Delta(\mathbf{y}_0, \mathbf{y}_1) + \frac{mL_\theta^2}{2\lambda}\Delta(\mathbf{y}_0, \mathbf{y}_1)^2, \tag{21}$$

where $\tilde{M} = Me^M$ with M the log-Lipschitz constant in Assumption 3.2 (i). The rate is therefore $\mathcal{O}(m\Delta + m\Delta^2)$.

Consider the case where $q(\hat{\mathbf{z}} \mid \mathbf{x}, \hat{\mathbf{y}})$ uses a multivariate Normal instead of Student-t

$$q(\hat{\mathbf{z}} \mid \mathbf{x}, \hat{\mathbf{y}} = \mathbf{y}_0) \coloneqq \mathcal{N}(\boldsymbol{\mu}_{\mathbf{y}_0}, \boldsymbol{\Sigma}_{\mathbf{y}_0}), \tag{22}$$
$$q(\hat{\mathbf{z}} \mid \mathbf{x}, \hat{\mathbf{y}} = \mathbf{y}_1) \coloneqq \mathcal{N}(\boldsymbol{\mu}_{\mathbf{y}_1}, \boldsymbol{\Sigma}_{\mathbf{y}_1}). \tag{23}$$

Proof of Theorem C.1. The KL divergence between two multivariate Normal distributions has a closed form

$$D_{\mathrm{KL}}[\mathcal{N}(\boldsymbol{\mu}_{\mathbf{y}_1}, \boldsymbol{\Sigma}_{\mathbf{y}_1}) \,||\, \mathcal{N}(\boldsymbol{\mu}_{\mathbf{y}_0}, \boldsymbol{\Sigma}_{\mathbf{y}_0})] = \frac{1}{2}\log\frac{\det(\boldsymbol{\Sigma}_{\mathbf{y}_0})}{\det(\boldsymbol{\Sigma}_{\mathbf{y}_1})} - \frac{m}{2} + \frac{1}{2}\operatorname{tr}(\boldsymbol{\Sigma}_{\mathbf{y}_0}^{-1}\boldsymbol{\Sigma}_{\mathbf{y}_1}) + \frac{1}{2}\operatorname{tr}\big(\boldsymbol{\Sigma}_{\mathbf{y}_0}^{-1}(\boldsymbol{\mu}_{\mathbf{y}_0} - \boldsymbol{\mu}_{\mathbf{y}_1})(\boldsymbol{\mu}_{\mathbf{y}_0} - \boldsymbol{\mu}_{\mathbf{y}_1})^\top\big).$$

Applying the determinant ratio bound (same as in Theorem B.4, with $\tilde{M} = Me^M$ from the assumption's log-Lipschitz constant) gives

$$\frac{1}{2}\log\frac{\det(\boldsymbol{\Sigma}_{\mathbf{y}_0})}{\det(\boldsymbol{\Sigma}_{\mathbf{y}_1})} \le \frac{\tilde{M}m}{2}\Delta(\mathbf{y}_0, \mathbf{y}_1). \tag{24}$$

The trace term follows the same approach as the Student-t case, using $\sigma^2_{\theta,j}(\mathbf{x}, \mathbf{y}_1)/\sigma^2_{\theta,j}(\mathbf{x}, \mathbf{y}_0) \le 1 + \tilde{M}\Delta$

$$\operatorname{tr}(\boldsymbol{\Sigma}_{\mathbf{y}_0}^{-1}\boldsymbol{\Sigma}_{\mathbf{y}_1}) = \sum_{j=1}^{m}\frac{\sigma^2_{\theta,j}(\mathbf{x}, \mathbf{y}_1)}{\sigma^2_{\theta,j}(\mathbf{x}, \mathbf{y}_0)} \le m(1 + \tilde{M}\Delta) = m + m\tilde{M}\Delta(\mathbf{y}_0, \mathbf{y}_1). \tag{25}$$

The constant m cancels with the $-m/2$ term in the KL decomposition. Under Assumption 3.2 (ii), the mean-shift term is bounded by

$$\operatorname{tr}\big(\boldsymbol{\Sigma}_{\mathbf{y}_0}^{-1}(\boldsymbol{\mu}_{\mathbf{y}_0} - \boldsymbol{\mu}_{\mathbf{y}_1})(\boldsymbol{\mu}_{\mathbf{y}_0} - \boldsymbol{\mu}_{\mathbf{y}_1})^\top\big) \le \frac{mL_\theta^2}{\lambda}\Delta(\mathbf{y}_0, \mathbf{y}_1)^2. \tag{26}$$

Consolidating all terms yields

$$\begin{aligned} D_{\mathrm{KL}}[q(\hat{\mathbf{z}} \mid \mathbf{x}, \hat{\mathbf{y}} = \mathbf{y}_1) \,||\, q(\hat{\mathbf{z}} \mid \mathbf{x}, \hat{\mathbf{y}} = \mathbf{y}_0)] &\le \frac{\tilde{M}m}{2}\Delta - \frac{m}{2} + \frac{1}{2}\big(m + m\tilde{M}\Delta\big) + \frac{mL_\theta^2}{2\lambda}\Delta^2 \\ &= \tilde{M}m\,\Delta(\mathbf{y}_0, \mathbf{y}_1) + \frac{mL_\theta^2}{2\lambda}\Delta(\mathbf{y}_0, \mathbf{y}_1)^2, \end{aligned} \tag{27}$$

where the $-m/2$ and $+m/2$ terms cancel and the two $\tilde{M}m\Delta/2$ terms sum to $\tilde{M}m\Delta$, and the bound on the mean-shift term uses $\|\boldsymbol{\Sigma}_{\mathbf{y}_0}^{-1}\|_2 \le 1/\lambda$ together with Assumption 3.2 (ii). This is the bound stated in Theorem C.1, and it is $\mathcal{O}\big(m\Delta + m\Delta^2\big)$. □

The mixed linear–quadratic rate contrasts with the at-most-linear rate $\mathcal{O}(\Delta)$ (with a label-independent intercept) under Student-t established in Theorem B.4. At small label discrepancies the Gaussian bound vanishes linearly ($\mathcal{O}(m\Delta)$), while the Student-t has a constant intercept C_1 (dominated by the $A\log S^2$ term); at large discrepancies the quadratic term $\mathcal{O}(m\Delta^2)$ can dominate the Gaussian bound when $L_\theta^2\Delta/(2\lambda) > \tilde{M}$; hence the two families differ in the functional form of their worst-case growth. This distinction motivates the heavy-tailed choice: the Student-t bound contains no quadratic term in Δ, at the cost of a fixed intercept, whereas the Gaussian bound carries one.

D Proof of Theorem 3.7

Proof. The total covariance error can be decomposed into three sources: (i) the label noise itself, captured by the difference between the true covariance $\boldsymbol{\Sigma}_{\mathbf{y}}$ and the observed noisy covariance $\boldsymbol{\Sigma}_{\hat{\mathbf{y}}}$; (ii) reconstruction error from sampling, $\boldsymbol{\Sigma}_{\hat{\mathbf{y}}} \to \boldsymbol{\Sigma}_{\hat{\mathbf{y}}_r}$; and (iii) the latent shift error, $\boldsymbol{\Sigma}_{\hat{\mathbf{y}}_r} \to \boldsymbol{\Sigma}_{\mathbf{y}_r}$, which arises because the decoder maps different latent points ($\hat{\mathbf{z}}$ vs. $\mathbf{z}$) to different Bernoulli parameters. We bound the total error via the triangle inequality

$$\mathbb{E}[\|\boldsymbol{\Sigma}_{\mathbf{y}} - \boldsymbol{\Sigma}_{\mathbf{y}_r}\|_F] \le \underbrace{\|\boldsymbol{\Sigma}_{\mathbf{y}} - \boldsymbol{\Sigma}_{\hat{\mathbf{y}}}\|_F}_{\epsilon_{\text{shift}}} + \underbrace{\mathbb{E}[\|\boldsymbol{\Sigma}_{\hat{\mathbf{y}}} - \boldsymbol{\Sigma}_{\hat{\mathbf{y}}_r}\|_F]}_{\epsilon_{\text{recon}}} + \mathbb{E}[\|\boldsymbol{\Sigma}_{\hat{\mathbf{y}}_r} - \boldsymbol{\Sigma}_{\mathbf{y}_r}\|_F]. \tag{28}$$

The terms ϵ_{shift} and ϵ_{recon} are intrinsic to the data and the sampling process, respectively; they are independent of the latent representation geometry. The geometric aspects of the problem are captured entirely by the third term, which we now bound.

Let $\mathbf{p} = g_\phi(\mathbf{x}, \mathbf{z})$ and $\mathbf{q} = g_\phi(\mathbf{x}, \hat{\mathbf{z}})$ denote the Bernoulli parameters decoded from the clean and noisy latent representations. For a multivariate Bernoulli random vector $\mathbf{y}_r \sim \mathrm{Bern}(\mathbf{p})$, the second moment decomposes elementwise as $\mathbb{E}[y_{r,i}y_{r,j}] = p_i p_j$ for $i \neq j$ and $\mathbb{E}[y_{r,i}^2] = \mathbb{E}[y_{r,i}] = p_i$ (since $y_{r,i}$ is binary). Assembling these into matrix form gives

$$\mathbb{E}[\mathbf{y}_r \mathbf{y}_r^\top] = \mathbf{p}\mathbf{p}^\top + \mathrm{diag}(\mathbf{p} - \mathbf{p} \circ \mathbf{p}), \tag{29}$$

$$\mathbb{E}[\hat{\mathbf{y}}_r \hat{\mathbf{y}}_r^\top] = \mathbf{q}\mathbf{q}^\top + \mathrm{diag}(\mathbf{q} - \mathbf{q} \circ \mathbf{q}), \tag{30}$$

where $\circ$ denotes elementwise multiplication.

Combining the triangle inequality with the Lipschitz continuity of g_ϕ gives

$$\begin{aligned}
\|\mathbf{\Sigma}_{\hat{\mathbf{y}}_r} - \mathbf{\Sigma}_{\mathbf{y}_r}\|_F &\leq \|\mathbf{q}\mathbf{q}^\top - \mathbf{p}\mathbf{p}^\top\|_F + \|\,\mathrm{diag}(\mathbf{q} - \mathbf{q} \circ \mathbf{q} - \mathbf{p} + \mathbf{p} \circ \mathbf{p})\|_F \\
&\leq \|\mathbf{q} + \mathbf{p}\|_2 \|\mathbf{q} - \mathbf{p}\|_2 + \|(\mathbf{q} - \mathbf{p}) \circ (\mathbf{1} - \mathbf{q} - \mathbf{p})\|_2 \\
&\leq \|\mathbf{q} + \mathbf{p}\|_2 \|\mathbf{q} - \mathbf{p}\|_2 + \|\mathbf{q} - \mathbf{p}\|_2 \cdot \|\mathbf{1} - \mathbf{q} - \mathbf{p}\|_\infty \\
&\leq (\|\mathbf{q} + \mathbf{p}\|_2 + 1)\|\mathbf{q} - \mathbf{p}\|_2 \\
&\leq L_\phi(\|\mathbf{q} + \mathbf{p}\|_2 + 1)\|\hat{\mathbf{z}} - \mathbf{z}\|_2,
\end{aligned} \tag{31}$$

where the second line uses the algebraic identity $(\mathbf{q} - \mathbf{q}\circ\mathbf{q}) - (\mathbf{p} - \mathbf{p}\circ\mathbf{p}) = (\mathbf{q} - \mathbf{p}) \circ (\mathbf{1} - \mathbf{q} - \mathbf{p})$, and the fourth line uses $\|\mathbf{1} - \mathbf{q} - \mathbf{p}\|_\infty \leq 1$ since $q_i, p_i \in [0, 1]$ implies $|1 - q_i - p_i| \leq 1$. This unified bound captures both the outer-product difference and the Bernoulli variance difference, obviating the need for a separate C_{bern} term.

To bound $\|\mathbf{q} + \mathbf{p}\|_2$, we exploit sparsity. Let $\mathcal{S}$ be the support of the d_ϵ largest probabilities

$$\|\mathbf{p}\|_1 = \sum_{i \in \mathcal{S}} p_i + \sum_{i \notin \mathcal{S}} p_i \leq d_\epsilon \pi_{\max} + \epsilon = \rho.$$

Using $d_w(\mathbf{p}) = \|\mathbf{p}\|_1^2 / \|\mathbf{p}\|_2^2 \geq d_{\min}$ (Assumption 3.6 (iv))

$$\|\mathbf{p}\|_2 \leq \frac{\|\mathbf{p}\|_1}{\sqrt{d_w(\mathbf{p})}} \leq \frac{\rho}{\sqrt{d_{\min}}}, \quad \|\mathbf{q}\|_2 \leq \frac{\rho}{\sqrt{d_{\min}}}.$$

Thus $\|\mathbf{q} + \mathbf{p}\|_2 \leq \|\mathbf{q}\|_2 + \|\mathbf{p}\|_2 \leq \frac{2\rho}{\sqrt{d_{\min}}}$.

Combining with Theorem D.1 yields

$$\mathbb{E}[\|\mathbf{\Sigma}_{\hat{\mathbf{y}}_r} - \mathbf{\Sigma}_{\mathbf{y}_r}\|_F] \leq \left(\frac{2\rho}{\sqrt{d_{\min}}} + 1\right) L_\phi \left(\sigma_{\mathbf{z}} + (L_\kappa + 1)\sigma_{\hat{\mathbf{z}}} + R\right) = C_{\mathrm{latent}},$$

which establishes the upper bound of Equation (9).

For the lower bound, we apply the reverse triangle inequality to the same three covariance matrices $\mathbf{\Sigma}_{\mathbf{y}}, \mathbf{\Sigma}_{\hat{\mathbf{y}}}, \mathbf{\Sigma}_{\mathbf{y}_r}$

$$\|\mathbf{\Sigma}_{\mathbf{y}} - \mathbf{\Sigma}_{\mathbf{y}_r}\|_F \geq \|\mathbf{\Sigma}_{\mathbf{y}} - \mathbf{\Sigma}_{\hat{\mathbf{y}}}\|_F - \|\mathbf{\Sigma}_{\hat{\mathbf{y}}} - \mathbf{\Sigma}_{\mathbf{y}_r}\|_F = \epsilon_{\mathrm{shift}} - \|\mathbf{\Sigma}_{\hat{\mathbf{y}}} - \mathbf{\Sigma}_{\mathbf{y}_r}\|_F. \tag{32}$$

Bounding the second term by the triangle inequality through $\mathbf{\Sigma}_{\hat{\mathbf{y}}_r}$ and taking expectations,

$$\begin{aligned}
\mathbb{E}\left[\|\mathbf{\Sigma}_{\mathbf{y}} - \mathbf{\Sigma}_{\mathbf{y}_r}\|_F\right] &\geq \epsilon_{\mathrm{shift}} - \mathbb{E}\left[\|\mathbf{\Sigma}_{\hat{\mathbf{y}}} - \mathbf{\Sigma}_{\hat{\mathbf{y}}_r}\|_F\right] - \mathbb{E}\left[\|\mathbf{\Sigma}_{\hat{\mathbf{y}}_r} - \mathbf{\Sigma}_{\mathbf{y}_r}\|_F\right] \\
&\geq \epsilon_{\mathrm{shift}} - \epsilon_{\mathrm{recon}} - C_{\mathrm{latent}},
\end{aligned} \tag{33}$$

where the second inequality uses the definitions of $\epsilon_{\mathrm{recon}}$ and the bound on $\mathbb{E}[\|\mathbf{\Sigma}_{\hat{\mathbf{y}}_r} - \mathbf{\Sigma}_{\mathbf{y}_r}\|_F]$ established above. Combining with the upper bound yields Equation (9). The lower bound is non-vacuous whenever $\epsilon_{\mathrm{shift}} > \epsilon_{\mathrm{recon}} + C_{\mathrm{latent}}$, i.e., when the intrinsic noise exceeds the recovery ability of the representation. □

Lemma D.1. *Under Assumptions 3.6 (i) and 3.6 (iii) we have*

$$\mathbb{E}[\|\mathbf{z} - \hat{\mathbf{z}}\|_2] \leq \sigma_{\mathbf{z}} + (L_\kappa + 1)\sigma_{\hat{\mathbf{z}}} + R, \tag{34}$$

where $\sigma_{\mathbf{z}} = \sqrt{\mathrm{tr}(\mathbf{\Sigma}_{\mathbf{z}})}$ *and* $\sigma_{\hat{\mathbf{z}}} = \sqrt{\frac{\nu}{\nu - 2}\,\mathrm{tr}(\mathbf{\Sigma}_{\hat{\mathbf{z}}})}$.

Proof. We decompose the Euclidean distance between clean and noisy latent representations into three components: the deviation of $\mathbf{z}$ from its own mean, the shift between the two means, and the deviation of $\hat{\mathbf{z}}$ from its mean. By the triangle inequality,

$$\|\mathbf{z} - \hat{\mathbf{z}}\|_2 \leq \|\mathbf{z} - \boldsymbol{\mu}_{\mathbf{z}}\|_2 + \|\boldsymbol{\mu}_{\mathbf{z}} - \boldsymbol{\mu}_{\hat{\mathbf{z}}}\|_2 + \|\boldsymbol{\mu}_{\hat{\mathbf{z}}} - \hat{\mathbf{z}}\|_2.$$

We bound each term in expectation separately.

We begin with the first term. The clean latent $\mathbf{z} \sim \mathcal{N}(\boldsymbol{\mu}_{\mathbf{z}}, \boldsymbol{\Sigma}_{\mathbf{z}})$ has sub-Gaussian tails. By Jensen's inequality, $\mathbb{E}[\|\mathbf{z} - \boldsymbol{\mu}_{\mathbf{z}}\|_2] \leq \sqrt{\mathbb{E}[\|\mathbf{z} - \boldsymbol{\mu}_{\mathbf{z}}\|_2^2]} = \sqrt{\mathrm{tr}(\boldsymbol{\Sigma}_{\mathbf{z}})} = \sigma_{\mathbf{z}}$, where $\boldsymbol{\Sigma}_{\mathbf{z}} = \mathrm{diag}(\sigma_\kappa^2(\hat{\mathbf{z}}))$ depends on the random $\hat{\mathbf{z}}$; taking the expectation over $\hat{\mathbf{z}}$ and applying Jensen inside-out gives $\mathbb{E}[\|\mathbf{z} - \boldsymbol{\mu}_{\mathbf{z}}\|_2] \leq \sqrt{\mathbb{E}_{\hat{\mathbf{z}}}[\mathrm{tr}(\boldsymbol{\Sigma}_{\mathbf{z}}(\hat{\mathbf{z}}))]} =: \sigma_{\mathbf{z}}$.

The second term follows from the mean shift connecting the two distributions through the correction path. Applying the Lipschitz property of μ_κ (Assumption 3.6 (i)) to the correction mean $\boldsymbol{\mu}_{\mathbf{z}} = \mathbb{E}_{q(\mathbf{z}|\hat{\mathbf{z}})}[\mathbf{z}] = \mu_\kappa(\hat{\mathbf{z}})$ and the noisy mean $\boldsymbol{\mu}_{\hat{\mathbf{z}}} = \mu_\theta(\mathbf{x}, \hat{\mathbf{y}})$, we obtain

$$\mathbb{E}[\|\boldsymbol{\mu}_{\mathbf{z}} - \boldsymbol{\mu}_{\hat{\mathbf{z}}}\|_2] \leq L_\kappa \, \mathbb{E}[\|\hat{\mathbf{z}} - \boldsymbol{\mu}_{\hat{\mathbf{z}}}\|_2] + R \leq L_\kappa \sigma_{\hat{\mathbf{z}}} + R,$$

which uses Assumption 3.6 (iii) (correction-path boundedness) together with $\mathbb{E}[\|\hat{\mathbf{z}} - \boldsymbol{\mu}_{\hat{\mathbf{z}}}\|_2] \leq \sigma_{\hat{\mathbf{z}}}$, obtained by the same Jensen argument applied to the Student-t distribution.

For the final term, the noisy latent $\hat{\mathbf{z}} \sim \mathrm{T}_\nu(\boldsymbol{\mu}_{\hat{\mathbf{z}}}, \boldsymbol{\Sigma}_{\hat{\mathbf{z}}})$ with $\nu > 2$ (ensuring finite second moments). The scaled dispersion satisfies $\mathbb{E}[\|\hat{\mathbf{z}} - \boldsymbol{\mu}_{\hat{\mathbf{z}}}\|_2^2] = \frac{\nu}{\nu-2} \mathrm{tr}(\boldsymbol{\Sigma}_{\hat{\mathbf{z}}}) = \sigma_{\hat{\mathbf{z}}}^2$, hence by Jensen $\mathbb{E}[\|\boldsymbol{\mu}_{\hat{\mathbf{z}}} - \hat{\mathbf{z}}\|_2] \leq \sigma_{\hat{\mathbf{z}}}$.

Summing the three expectations yields the stated bound. □

E Datasets, Implementation, and Sensitivity

Table 2: Dataset statistics.

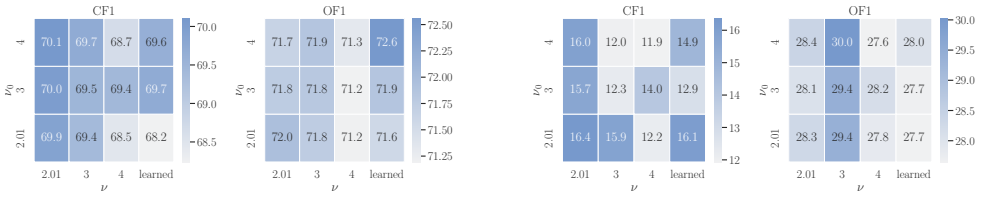
Dataset	Training	Validation	Clean	Test	Labels
VOC07	4,509	251	251	4,952	20
VOC12	10,386	577	577	4,952	20
COCO	82,081	10,034	10,034	20,069	80
Tomato	3,945	562	562	584	8
DeepFashion	14,000	1,000	1,000	4,000	26

We use a β-VAE with $\beta = 0.001$ (the effective weight under mean-reduced reconstruction), $\nu = \nu_0 = 2.001$, $\eta = 0.1$ (see Section E.3), latent dimension $m = 32$, and label-embedding dimension 32. The small β is deliberate; it keeps the KL regularization weak such that reconstruction fidelity dominates, which the confidence-sharpening analysis in Section G and the η sensitivity study suggest is the favorable operating region. We do not sweep β here, since it trades reconstruction fidelity against latent regularity. The label encoder is a multi-layer MLP with GELU activation and Batch Normalization (64-unit layers). The VAE is trained with SGD (learning rate 0.1, weight decay 10^{-3}) with a cosine schedule, batch size 512 (scaled down on a single GPU), for 15 epochs on VOC07 (unsupervised; 5 epochs in the semi-supervised stage) and analogous budgets on the other datasets. Base models use SGD or AdamW with learning rates in $[10^{-5}, 2\times10^{-4}]$ depending on the classifier. All hyperparameters are consistent across settings, and the ablation runs reuse this protocol exactly.

Sensitivity analysis confirms stable performance across ν values; the role of η is examined in Section E.3.

E.1 Sensitivity of the Degrees of Freedom

Figure 3 presents the sensitivity analysis over ν_0 and ν. On VOC07 the base classifier is HLC_r under Pair-40% noise; on Tomato it is $\mathrm{MLP}_{\mathrm{lev}}$ under Sym-40% noise, where $\mathrm{MLP}_{\mathrm{lev}}$ denotes a standard MLP head over the LeViT (Graham et al., 2021) encoder, distinct from the ViT-224 subscript v used elsewhere. The grid covers $\nu_0 \in \{1, 2.01, 3, 4\}$ and $\nu \in \{2.01, 3, 4\}$, together with a *learned* variant in which the degrees of freedom are inferred from the input; each cell averages five repetitions.



(a) VOC07 (HLC_r, Pair-40%) (b) Tomato (MLP_{lev}, Sym-40%)

Figure 3: Sensitivity of LSNPC to the degrees of freedom ν_0 and ν. Each value averages five repetitions. The *learned* setting infers ν from the input instead of fixing it.

Values just above 2 follow the requirement of Theorem 3.3 that $\nu > 2$; the main experiments use $\nu = \nu_0 = 2.001$, whose role on the grid is played by the near-boundary representative 2.01 (shown at two-decimal resolution).

Under the learned setting, the model infers the degrees of freedom directly from $(\mathbf{x}, \hat{\mathbf{y}})$. The variational distribution becomes

$$q(\hat{\mathbf{z}} \mid \mathbf{x}, \hat{\mathbf{y}}) = \mathrm{T}_{\nu_\theta(\mathbf{x},\hat{\mathbf{y}})}\big(\mu_\theta(\mathbf{x}, \hat{\mathbf{y}}), \mathrm{diag}(\sigma^2_\theta(\mathbf{x}, \hat{\mathbf{y}}))\big)\,,$$

with

$$\nu_\theta(\mathbf{x}, \hat{\mathbf{y}}) = \mathrm{ReLU}(\mathrm{Net}(\mathbf{x}, \hat{\mathbf{y}})) + 1,$$

where $\mathrm{ReLU}(\cdot)$ ensures non-negativity and the addition of 1 keeps the degrees of freedom strictly above the minimum permissible value for the Student-t distribution.

On VOC07, performance is generally robust to the choices of ν_0 and ν across all three metrics, and the weakest configurations degrade only marginally. A consistent trend is that performance decreases when $\nu > \nu_0$. The learned ν sometimes attains the best OF1, whereas the fixed near-boundary value yields the strongest and most stable CF1 while keeping OF1 competitive. This suggests that a learnable ν biases the model toward majority labels, whereas a fixed heavy-tailed setting regularizes the correction to also preserve minority-label signal; moreover, when ν is learnable, the choice of ν_0 becomes more sensitive.

On Tomato, the learned ν exhibits more pronounced variability, particularly in CF1. Although $\nu = 3$ attains the highest OF1, its CF1 is substantially lower in most configurations. The pair $(\nu_0, \nu) = (2.001, 2.001)$ is therefore adopted in all main experiments as a stable, consistently strong default.

E.2 Semi-Supervised Training Schedule

The semi-supervised objective of Section 2 is optimized with the two-phase schedule in Algorithm 1. Within each epoch, an unsupervised pass over the noisy dataset is followed by a supervised pass over the clean subset. This staged schedule lets the clean labels guide the correction path without destabilizing the unsupervised objective.

Algorithm 1: Semi-supervised training schedule

Input: learning rate ξ
while *not converged* **do**
  Unsupervised pass over $\tilde{\mathcal{D}}$;
  foreach *batch* $\tilde{\mathcal{B}} \subset \tilde{\mathcal{D}}$ **do**
    $\Phi \leftarrow \Phi - \xi \nabla_\Phi \mathcal{U}(\tilde{\mathcal{B}})$;
  Supervised pass over the clean subset $\mathcal{D}$;
  foreach *batch* $\mathcal{B} \subset \mathcal{D}$ **do**
    $\Phi \leftarrow \Phi - \xi \nabla_\Phi \mathcal{L}(\mathcal{B})$;

E.3 Hyperparameter Sensitivity of the Mixing Coefficient η

The semi-supervised posterior is a mixture $q(\mathbf{z} \mid \mathbf{x}, \mathbf{y}, \hat{\mathbf{z}}) = \eta\, q(\mathbf{z} \mid \mathbf{x}, \mathbf{y}) + (1-\eta)\, q(\mathbf{z} \mid \hat{\mathbf{z}})$, where the η branch encodes $\mathbf{z}$ directly from clean labels and the $(1-\eta)$ branch encodes $\mathbf{z}$ from the noisy latent $\hat{\mathbf{z}}$ via the correction path $q(\mathbf{z} \mid \hat{\mathbf{z}})$. The parameter η controls how often the model routes through the clean-label encoder during training. To examine the effect of η, we evaluate the semi-supervised setting on VOC12 across two configurations: ASL (ViT-224, Pair-30% noise, 5% clean subset) and HLC (ResNet-50, Sym-50% noise, 5% clean subset). We sweep $\eta \in \{0.0, 0.25, 0.5, 0.75, 1.0\}$ with three to five random repetitions per setting.

Table 3: Sensitivity to the mixing coefficient η on VOC2012. Mean $\pm$ SE over independent repetitions.

Dataset	η	CF1↑	OF1↑	mAP↑
VOC2012 ASL, ViT-224 Pair-30%	0.00	88.60 ± 0.25	$\mathbf{89.03} \pm 0.26$	$\mathbf{93.23} \pm 0.07$
	0.25	$\mathbf{88.62} \pm 0.23$	$\mathbf{89.03} \pm 0.16$	$\mathbf{93.23} \pm 0.07$
	0.50	88.39 ± 0.19	88.58 ± 0.13	92.91 ± 0.15
	0.75	88.05 ± 0.59	88.53 ± 0.34	92.99 ± 0.07
	1.00	87.91 ± 0.39	88.50 ± 0.37	92.76 ± 0.13
VOC2012 HLC, RN50 Sym-50%	0.00	$\mathbf{81.35} \pm 0.51$	$\mathbf{83.19} \pm 0.33$	90.40 ± 0.15
	0.25	81.03 ± 0.65	82.27 ± 0.38	$\mathbf{90.50} \pm 0.10$
	0.50	80.43 ± 0.62	82.52 ± 0.14	90.46 ± 0.11
	0.75	80.59 ± 0.84	81.90 ± 0.49	90.24 ± 0.13
	1.00	80.82 ± 0.79	80.34 ± 2.57	89.85 ± 0.22

Table 3 shows a consistent pattern; performance is maximized at $\eta = 0$ or $\eta = 0.25$, and degrades as η increases beyond 0.5. Setting $\eta = 1.0$, which routes exclusively through the clean-label encoder $q(\mathbf{z} \mid \mathbf{x}, \mathbf{y})$ during training, produces the worst results overall. This pattern holds across both encoder architectures (convolutional and transformer) and both noise types (pairflip and symmetric). This result may appear counterintuitive. The clean-label encoder branch is explicitly designed to exploit clean data, yet ignoring it entirely ($\eta = 0$) yields equal or superior performance. The explanation lies in how clean labels reach the $\mathbf{z}$-encoding pathway.

During inference, correction proceeds by encoding $\hat{\mathbf{z}} \sim q(\hat{\mathbf{z}} \mid \mathbf{x}, \hat{\mathbf{y}})$ from the noisy prediction, then sampling $\mathbf{z} \sim q(\mathbf{z} \mid \hat{\mathbf{z}})$ and decoding $\mathbf{y} \sim p(\mathbf{y} \mid \mathbf{x}, \mathbf{z})$. The clean-label encoder $q(\mathbf{z} \mid \mathbf{x}, \mathbf{y})$ is never evaluated at test time. When $\eta > 0$, the model allocates a portion of its capacity to a $\mathbf{z}$-encoding pathway exercised only during training. This creates a distribution mismatch. The decoder g_ϕ sees $\mathbf{z}$ drawn from two different distributions during training (one from clean labels and one from noisy $\hat{\mathbf{z}}$); it must generalize solely from the noisy-$\hat{\mathbf{z}}$ distribution at test time. As η increases, less gradient signal reaches the inference-relevant network $q(\mathbf{z} \mid \hat{\mathbf{z}})$.

Clean labels under $\eta = 0$ still guide the $\hat{\mathbf{z}} \to \mathbf{z}$ mapping through the reconstruction loss. The supervised loss $\mathrm{BCE}\big(g_\phi(\mathbf{x}, \mathbf{z}), \mathbf{y}\big) + \mathrm{BCE}\big(g_\phi(\mathbf{x}, \hat{\mathbf{z}}), \hat{\mathbf{y}}\big)$ is computed on every supervised batch regardless of η. The first term compares the decoder's output from $\mathbf{z}$ (which at $\eta = 0$ is always $\mathbf{z} \sim q(\mathbf{z} \mid \hat{\mathbf{z}})$) against the true labels $\mathbf{y}$. The resulting gradient propagates through the decoder g_ϕ into $\mathbf{z}$, then through $q(\mathbf{z} \mid \hat{\mathbf{z}})$ and further into $\hat{\mathbf{z}}$ and $q(\hat{\mathbf{z}} \mid \mathbf{x}, \hat{\mathbf{y}})$, providing a direct learning signal for the correction-path encoder $q(\mathbf{z} \mid \hat{\mathbf{z}})$. Clean labels therefore do not need a separate encoder branch to refine the latent representation; the shared decoder and reconstruction loss already couple the clean-label objective to the $\hat{\mathbf{z}} \to \mathbf{z}$ mapping through the gradient path. The mixing parameter η controls only which $\mathbf{z}$-encoding pathway is regularized by the KL divergence to the prior; it does not control whether clean labels inform the correction mechanism.

This analysis suggests that any small value of the mixing parameter ($\eta \leq 0.25$) yields comparable performance, with no need for a finely tuned optimum; we set $\eta = 0.1$ in the main experiments. At $\eta = 0$ the semi-supervised objective simplifies, with the $\mathbf{z}$-path always following $q(\mathbf{z} \mid \hat{\mathbf{z}})$ (the inference-time correction path); clean labels contribute through the reconstruction loss gradient that reaches $q(\mathbf{z} \mid \hat{\mathbf{z}})$ via $\mathbf{z}$, without requiring an auxiliary encoder branch.

F EXTENSION OF NPC TO MULTILABEL

Originally designed for multiclass classification, NPC (Bae et al., 2022) minimizes a reconstruction loss with two KL terms: a latent-space regularizer $D_{\mathrm{KL}}(q(\mathbf{z} \mid \mathbf{x}, \hat{\mathbf{y}}) \mid\mid p(\mathbf{z}))$ and a label-space consistency term $D_{\mathrm{KL}}(q(\mathbf{y}_{\text{recon}} \mid \mathbf{x}, \hat{\mathbf{y}}) \mid\mid p_h(\hat{\mathbf{y}} \mid \mathbf{x}))$. We extend NPC to multilabel (NPC_{ext}) by instantiating per-label Bernoulli outputs; in this setting, $\mathbf{y}$ and $\hat{\mathbf{y}}$ are multivariate Bernoulli, and the label-space KL decomposes as a sum of k independent per-label terms

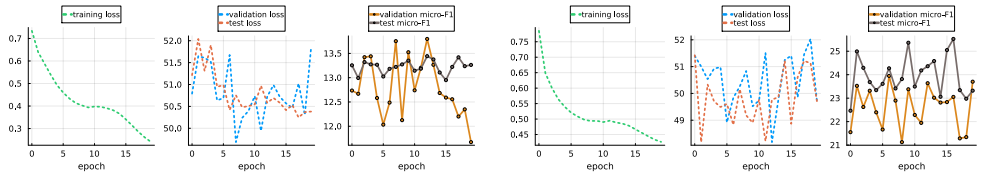
$$D_{\mathrm{KL}}(P \mid\mid Q) = \sum_{i=1}^{k} \left[p_i \log \frac{p_i}{q_i} + (1 - p_i) \log \frac{1 - p_i}{1 - q_i} \right]. \tag{35}$$

NPC_{ext} converges to outputs that are independent of noise type, noise rate, and base classifier, and sit at the level of a trivial majority-class baseline. On datasets with larger label spaces and lower positive rates the fixed point is correspondingly lower; on smaller label spaces with denser positive rates it sits higher. In every case the behavior is noise-invariant, which is the defining signature of the cascade below.

This failure under multilabel sparsity proceeds through three interacting mechanisms. First, sparse multilabel targets allow the decoder to achieve a low reconstruction loss by outputting dataset-level Bernoulli marginals without using the latent variable $\mathbf{z}$. The encoder is then starved of gradient and collapses; $D_{\mathrm{KL}}(q(\mathbf{z} \mid \mathbf{x}, \hat{\mathbf{y}}) \mid\mid p(\mathbf{z})) \to 0$, driving $\mathbf{z}$ toward the prior for all inputs. Once collapsed, $\mathbf{z}$ carries no instance-specific information, hence instance-level correction is structurally impossible regardless of subsequent gradient signals; this alone makes the output invariant to the input. Second, with $\mathbf{z}$ collapsed, the label-space KL alone determines the per-label outputs, and under multilabel sparsity the aggregate KL weight on the abundant negative-label terms vastly exceeds that on the rare positives, swamping the reconstruction signal from the latter. The optimizer therefore converges to a fixed point in which each per-label prediction equals the dataset's marginal positive rate, kept near the classifier's average behavior by the label-space KL, precisely the behavior of a trivial marginal-probability baseline. Third, even if posterior collapse were prevented, sparsity would still dilute the correction gradient: in a typical batch the overwhelming majority of per-label predictions are correctly predicted negatives; hence the gradient from the rare correction-worthy positives is diluted relative to the preserving gradient from the negatives. The optimizer reaches a flat basin in which the expected gradient is near zero, which is why longer training, larger batches, or learning-rate tuning cannot escape the fixed point.

The softmax normalization $\sum_c p_c = 1$ couples all class predictions per instance; hence correcting one class's probability forces the others to adjust; this prevents encoder collapse and gives the label-space KL balanced influence on every prediction. The multilabel extension uses independent sigmoids, which decouple the labels and remove this structural safeguard. With decoupled labels, sparsity triggers dilution, which triggers collapse, which in turn locks the predictions onto the marginal fixed point, a cascade absent in the coupled multiclass setting.

We validated this analysis with a Normal copula implementation (Nelsen, 2006; Wang & Yin, 2020) to capture label correlations; Figure 4 shows the loss plateauing at the marginal fixed point. The convergence is stable rather than random, ruling out a "random wandering" interpretation of the behavior.



(a) VOC07, MLP_r, Pair-30% (b) Tomato, ADDGCN, Sym-50%

Figure 4: Training analysis of the multilabel NPC extension.

G Detailed mAP Analysis

Table 1 reveals a consistent pattern; LSNPC improves CF1 and OF1 in most settings, while mAP, a ranking metric, behaves more erratically. CF1 and OF1 are threshold metrics, indicating that any correction that pushes a label probability across $p \geq 0.5$ improves them, regardless of rank ordering. mAP, in contrast, averages per-class average precision (AP), sorting test instances by confidence within each class and integrating the precision–recall curve; swapping the scores of two true positives degrades AP even if both remain correctly classified at the fixed threshold (Wu & Zhou, 2017).

The ELBO provides no ranking signal; its reconstruction and KL-regularization terms operate per-instance and per-label, and Theorem 3.7 bounds the expected Frobenius error between the true and reconstructed label covariances (LCN). Each covariance entry $\Sigma_{ij} = \frac{1}{N}\sum_n (y_{ni} - \mu_i)(y_{nj} - \mu_j)$ is a commutative sum over N instances; hence permuting the confidence scores within a class leaves it unchanged. Given accurate marginal Bernoulli parameters, the loss is identical for a correct ranking and a random permutation.

This invariance manifests as *confidence sharpening*: the Bernoulli decoder with weak KL regularization ($\beta = 0.001$) drives outputs toward 0 or 1 (Guo et al., 2017). Crossing the 0.5 threshold improves CF1 and OF1, while concentrated outputs scramble within-class ordering. If four true positives with base scores $\{0.48, 0.45, 0.42, 0.38\}$ (below threshold, yet correctly ranked) sharpen to $\{0.92, 0.04, 0.95, 0.08\}$, CF1 improves because two instances now exceed 0.5, while the ranking among positives is randomized; hence mAP degrades.

The effect depends on noise type and base classifier. Under moderate symmetric noise (30–40%), noise-aware baselines (MCM, HLC, ADDGCN) already learn effective within-class rankings, and LSNPC's post-hoc correction overrides this structure, dropping mAP. Under severe pairflip noise, the base ranking is systematically corrupted (entire label pairs are confused), and the learned clean–noisy correspondence restores ordering, yielding gains. Over the VOC07 settings in Table 1, symmetric noise leaves Δ_{mAP} near zero at 30–40% (worst: ASL_s at Sym-50%), while pairflip noise improves it by a few points (largest for MCM_s at Pair-40%, $+6.6$, where the base ranking was already poor). Semi-supervised results (Figure 2) corroborate this: even a small clean subset yields a more ranking-preserving latent geometry and mAP gains in most settings.

Two conclusions follow. First, LSNPC is most impactful under high noise or weak base classifiers; for strong noise-aware classifiers under mild symmetric noise, the base predictions may already be better for ranking-sensitive applications. Second, the threshold-versus-ranking gap points to a concrete extension: augmenting the ELBO with ranking-aware terms (e.g., listwise losses that penalize within-class rank inversions) could close the gap while retaining reconstruction fidelity, analogous to the pointwise-versus-listwise distinction in learning-to-rank.

H Additional Results

H.1 Additional Unsupervised Results

Table 4 reports results on Tomato; the VOC07 results appear in Table 1 in the main text. The pattern is consistent across datasets; LSNPC improves CF1 and OF1 across base classifiers and noise-aware baselines (UNM, ASL, LRAD), with the largest gains at 50% noise.

H.2 Sparsity Ablation

We measure representation quality under varying sparsity using the Label Covariance Norm (LCN), defined as the Frobenius norm of the difference between the covariance matrices of predicted and true label vectors

$$\text{LCN} = \|\boldsymbol{\Sigma}_{\mathbf{y}_r} - \boldsymbol{\Sigma}_{\mathbf{y}}\|_F\,, \tag{36}$$

where $\boldsymbol{\Sigma}_{\mathbf{y}}$ and $\boldsymbol{\Sigma}_{\mathbf{y}_r}$ denote the uncentered covariance matrices of true and predicted labels, respectively. We evaluate on VOC12 (MLP_r) under Sym-50% noise and on Tomato (ASL, ViT-224) under Pair-40% noise, partitioned by label density (Very Sparse: 1, Sparse: 2–3, Dense: 4–5 positives; on the 8-label Tomato dataset the Dense bin has no samples and is reported as N/A).

Table 4: Unsupervised correction on Tomato. **Bold**: LSNPC improves over baseline. NR: noise rate (%). Mean ± SE.

NR		Symmetric			Pairflip		
		CF1	OF1	mAP	CF1	OF1	mAP
30%	MLP_r	60.4	67.0	74.7	38.8	56.3	57.9
	+LSNPC	**63.7**±0.4	**72.3**±0.4	70.4±1.4	**45.7**±4.9	**59.3**±2.9	57.8±1.3
	UNM_r	32.3	52.4	58.6	41.9	45.5	62.1
	+LSNPC	**50.0**±0.9	**59.3**±1.3	56.3±1.1	**45.3**±2.1	**47.8**±2.8	**63.2**±0.9
	MLP_s	76.3	81.0	85.2	66.7	77.6	74.6
	+LSNPC	**80.0**±0.6	**84.5**±0.6	**85.7**±0.5	**70.6**±0.7	**80.0**±0.4	**75.0**±1.0
	MCM_s	74.0	78.9	80.0	52.3	70.4	73.0
	+LSNPC	63.8±1.2	71.8±1.6	**81.5**±1.5	**54.3**±2.3	62.0±2.7	71.0±1.6
	ASL_s	80.9	84.4	85.9	71.6	78.7	82.6
	+LSNPC	77.2±0.3	83.6±0.1	83.9±1.3	**74.1**±4.3	**80.9**±3.0	**83.2**±2.2
	$LRAD_s$	75.0	80.2	85.0	48.3	45.0	63.4
	+LSNPC	**79.2**±0.7	**83.8**±0.3	**86.6**±1.3	**50.4**±1.8	**55.6**±2.3	**71.2**±1.2
40%	MLP_r	40.1	57.9	67.7	33.3	49.9	53.4
	+LSNPC	**43.2**±1.6	**61.6**±0.5	65.5±0.9	**39.9**±2.1	**54.4**±0.7	**53.7**±0.4
	UNM_r	25.9	42.3	53.2	28.7	46.9	53.1
	+LSNPC	**46.2**±1.2	**56.8**±2.7	49.1±1.4	**45.9**±2.2	**51.6**±2.8	49.6±0.6
	MLP_s	62.6	66.8	79.0	52.3	61.0	65.2
	+LSNPC	**66.1**±3.2	**73.9**±1.6	**79.9**±2.1	**62.5**±2.0	**69.1**±1.3	**66.6**±0.8
	MCM_s	56.2	57.9	73.5	45.2	58.9	58.4
	+LSNPC	**57.4**±1.4	**61.9**±2.5	**73.7**±1.3	**52.0**±2.5	**59.1**±2.9	**60.7**±1.9
	ASL_s	66.8	72.5	79.4	64.8	73.0	70.4
	+LSNPC	**70.1**±1.1	**78.4**±0.9	75.9±1.5	62.4±2.2	**73.1**±1.1	68.3±0.3
	$LRAD_s$	71.3	72.0	81.9	18.7	23.5	48.1
	+LSNPC	**73.0**±1.6	**76.9**±2.0	81.8±1.0	**43.3**±3.2	**47.5**±4.1	**55.3**±1.9
50%	MLP_r	26.7	40.9	60.8	25.9	43.0	47.4
	+LSNPC	**30.7**±1.0	**45.8**±1.9	**61.2**±0.7	**31.2**±1.7	**45.7**±0.8	47.1±1.1
	UNM_r	10.2	17.8	49.6	21.3	38.9	45.8
	+LSNPC	**41.7**±1.1	**53.4**±1.1	43.2±1.7	**43.3**±2.4	**48.7**±2.5	43.3±1.8
	MLP_s	46.9	56.2	68.4	31.3	38.9	55.0
	+LSNPC	**53.9**±4.4	**65.7**±2.8	**70.6**±0.7	**37.9**±1.6	**44.1**±3.1	55.0±1.1
	MCM_s	34.1	46.5	60.0	31.7	41.8	48.5
	+LSNPC	**56.7**±1.1	**67.1**±1.8	**63.2**±0.4	**48.5**±0.6	**54.0**±0.6	**50.0**±1.1
	ASL_s	47.6	63.9	72.1	53.5	60.1	58.2
	+LSNPC	**60.6**±2.2	**73.2**±1.2	67.1±2.0	52.7±3.7	58.4±0.5	**62.5**±2.4
	$LRAD_s$	52.5	62.2	69.9	19.3	15.3	42.8
	+LSNPC	**59.2**±1.9	**70.5**±1.3	**74.4**±2.3	**37.8**±1.5	**39.3**±1.5	**48.9**±1.2

Table 5: Sparsity ablation. **Bold**: LSNPC improves over the base classifier. Instances are partitioned by label density (number of positive labels): Very Sparse (1), Sparse (2–3), Dense (4–5); Overall aggregates the full test set. VOC12: MLP_r, Sym-50%; Tomato: ASL, ViT-224, Pair-40%. N/A: no samples in this density bin on the 8-label Tomato dataset.

Configuration			Metrics			
Dataset	**Sparsity**	**Model**	**OF1↑**	**CF1↑**	**mAP↑**	**LCN↓**
VOC12 (MLP_r)	Very Sparse	Base	67.34	63.48	89.55	0.3567
		+LSNPC	**91.10**	**86.47**	**93.90**	**0.1805**
	Sparse	Base	53.98	43.54	80.20	0.5063
		+LSNPC	**76.65**	**68.30**	**84.92**	**0.4405**
	Dense	Base	42.06	22.07	65.02	0.7476
		+LSNPC	**60.47**	**33.18**	**72.98**	**0.7182**
	Overall	Base	59.05	55.08	86.96	0.4352
		+LSNPC	**82.22**	**78.46**	**90.23**	**0.3147**
Tomato (ASL_v)	Very Sparse	Base	66.52	56.60	65.95	0.3399
		+LSNPC	**67.97**	**60.58**	**67.99**	0.5086
	Sparse	Base	64.47	29.90	43.19	0.5776
		+LSNPC	**68.68**	**33.38**	**43.88**	0.6859
	Dense	Base	–	–	–	N/A
		+LSNPC	–	–	–	N/A
	Overall	Base	65.83	53.74	63.69	0.3703
		+LSNPC	**68.20**	**58.06**	**64.49**	0.5233

Table 5 reports the per-bin metrics. On VOC12 (MLP_r, Sym-50% noise), LSNPC improves every classification metric in every density bin and reduces LCN throughout, with the largest reductions on the sparsest instances, supporting Theorem 3.7; the learned latent geometry most effectively preserves label covariance on minority-label instances. Consistent with the two-sided characterization, the corrected LCN remains above the intrinsic noise floor in every bin while approaching it most closely on the sparsest instances. On Tomato (ASL, ViT-224, Pair-40% noise), the correction again improves all classification metrics; it nonetheless increases LCN. The Dense bin has no samples on the 8-label Tomato dataset (N/A). The Tomato result reflects the small label space: ASL's base predictions already approximate the true label covariance, and LSNPC's correction, which sharpens probabilities toward 0 or 1, produces a corrected covariance that deviates further from the base predictions', trading covariance fidelity for improved threshold-based decisions (see Section 4.2).

H.3 Student-t vs. Gaussian Representation

Table 6 tests the geometric prediction of Theorem 3.3 by replacing the Student-t shift with a Gaussian one (equal mean, rescaled variance). On VOC12 (ADDGCN, Sym), the advantage of the Student-t representation is large and grows with noise severity; CF1 79.1 vs. 70.6 at 30% and 70.7 vs. 49.4 at 50%, with a matching OF1 gap. The Student-t correction also recovers a substantial fraction of the base's loss at 50% ($50.7 \rightarrow 70.7$), whereas the Gaussian variant stays near the base, confirming that the heavy tail, not the variance rescaling alone, enables correction at severe noise. On VOC07 (MLP_{lev} (Graham et al., 2021), Pair), the two representations are closer, and at 50% both collapse toward the base as the base classifier itself degrades sharply under pairflip corruption. The contrast is consistent with the theory; the tail advantage is most pronounced precisely where label discrepancy is large and the Gaussian alternative saturates, and is muted in regimes where the base prediction is already poor for both variants.

Table 6: Student-t vs. Gaussian representation.

NR		VOC07, Pair, MLP_{lev}		VOC12, Sym, ADDGCN	
		CF1	OF1	CF1	OF1
30%	Baseline	65.5	71.3	74.1	74.5
	GAUSS	64.0 ± 3.1	72.2 ± 2.0	70.6 ± 2.1	72.6 ± 1.0
	LSNPC	$\mathbf{66.5 \pm 3.0}$	$\mathbf{73.5 \pm 1.7}$	$\mathbf{79.1 \pm 0.3}$	$\mathbf{80.4 \pm 0.1}$
50%	Baseline	16.4	27.4	50.7	59.0
	GAUSS	$\mathbf{16.4 \pm 1.9}$	30.6 ± 2.1	49.4 ± 0.4	53.0 ± 0.5
	LSNPC	16.1 ± 2.0	$\mathbf{33.0 \pm 1.9}$	$\mathbf{70.7 \pm 0.6}$	$\mathbf{76.8 \pm 0.3}$

H.4 Direct-$\hat{\mathbf{z}}$ Decoding

This ablation removes the correction step. Given a noisy prediction $\hat{\mathbf{y}}$, the encoder produces $\hat{\mathbf{z}} \sim q(\hat{\mathbf{z}} \mid \mathbf{x}, \hat{\mathbf{y}})$; instead of sampling $\mathbf{z} \sim q(\mathbf{z} \mid \hat{\mathbf{z}})$ and decoding from $\mathbf{z}$, we decode the labels directly from $\hat{\mathbf{z}}$, i.e. $\mathbf{y}^* \sim \text{Bern}(g_\phi(\mathbf{x}, \hat{\mathbf{z}}))$. If the latent displacement rather than the shared decoder carried most of the correction, direct-$\hat{\mathbf{z}}$ decoding would recover no more than the base classifier's own quality; if the shared decoder is the dominant pathway, it should already capture most of the gain, with the full LSNPC pipeline adding a smaller refinement. We report CF1, OF1, and mAP for (i) the base classifier, (ii) direct-$\hat{\mathbf{z}}$ decoding, and (iii) full LSNPC correction; the three arms share the same trained checkpoint and differ only in the prediction mode. Table 7 reports results on VOC12 (ResNet-50) for HLC_r (five repetitions) and MLP_r and UNM_r (three repetitions) at pairflip and symmetric noise of 30% and 50%. The left group is the unsupervised model; the right group repeats the same arms for the semi-supervised model (three repetitions, $\eta = 0.1$), which additionally trains on a small clean subset (Section 4.2).

Across all three classifiers and both noise types, direct-$\hat{\mathbf{z}}$ decoding captures most of the correction, and the explicit shift adds only a small increment. The shift's gain is largest under severe symmetric corruption, where the Student-t bound's linear rate matters most, and negligible under pairflip noise or at 30% noise, where both arms track the strong base. This mirrors the distributional division in Section 2; the heavy-tailed $q(\hat{\mathbf{z}} \mid \mathbf{x}, \hat{\mathbf{y}})$ already concentrates on the label-consistent region of the shared manifold, hence decoding $\hat{\mathbf{z}}$ directly corrects most of the error and the Gaussian $q(\mathbf{z} \mid \hat{\mathbf{z}})$ only recenters the residual. Consistent with Section 4.2, mAP can slightly degrade under strong symmetric noise even as CF1 and OF1 improve. Because all arms share one trained checkpoint, the comparison isolates the shift from checkpoint variation; for the MLP_r and UNM_r pairflip-50% rows, one of three repetitions was discarded (its corrected checkpoint fell 5–6 CF1 points below the others, a bf16 training artifact), leaving $n = 2$.

The semi-supervised group sharpens this decomposition. Direct-$\hat{\mathbf{z}}$ decoding captures essentially the entire correction, recovering 88–100% of the total gain across all twelve configurations, and the residual shift contributes at most 0.7 CF1 points, down from about 1.3 in the unsupervised model. The clean subset strengthens the shared-manifold alignment, placing $\hat{\mathbf{z}}$ closer to the clean manifold and leaving less displacement for the shift to bridge. It also removes the mAP caveat, with mAP improving over the base in every configuration, consistent with a sharper manifold producing well-ordered scores rather than confidence-sharpened ones.

H.5 Separate-Decoder Variant

The shared decoder is the architectural claim behind manifold alignment: it forces clean and noisy labels through the same decoding pathway. This ablation replaces g_ϕ with two decoders, g_ϕ^{clean} (decoding $\mathbf{y}$ from $\mathbf{z}$) and g_ϕ^{noisy} (decoding $\hat{\mathbf{y}}$ from $\hat{\mathbf{z}}$), and trains the same objectives otherwise unchanged.

We report the separate-decoder variant on the VOC07 (MLP_r, ResNet-50) configurations at 50% noise, five repetitions, alongside the shared-decoder default (the five-repetition runs of Table 1); per-repetition distributions are shown in Figure 5. The separate-decoder variant collapses to near-random performance (CF1 $\approx$ 11.2, OF1 $\approx$ 14.1, mAP $\approx$ 8.9), whereas the shared decoder at-

Table 7: Direct-$\hat{\mathbf{z}}$ decoding ablation on VOC12 (ResNet-50). Mean ± SE. **Bold**: best-performing arm per row within each group (ties bolded).

Config	Noise	NR	Unsupervised			Semi-supervised		Metric
			Base	Direct-$\hat{z}$	Full	Direct-$\hat{z}$	Full	
HLC_r	pairflip	30%	83.8	83.7±0.1	**83.9±0.1**	**86.6±0.1**	**86.6±0.1**	CF1
			85.4	85.0±0.2	**85.1±0.2**	87.9±0.0	**88.0±0.0**	OF1
			91.6	91.0±0.0	**91.1±0.0**	92.2±0.0	**92.3±0.0**	mAP
		50%	42.0	45.0±0.3	**45.5±0.3**	76.5±0.5	**76.6±0.5**	CF1
			47.2	48.8±0.4	**49.2±0.4**	77.4±0.3	**77.5±0.3**	OF1
			52.4	54.4±0.1	**54.5±0.1**	85.7±0.3	**86.0±0.3**	mAP
	symmetric	30%	85.9	85.5±0.1	**85.7±0.1**	**86.5±0.1**	**86.5±0.1**	CF1
			86.5	86.0±0.0	**86.1±0.1**	**87.2±0.0**	**87.2±0.1**	OF1
			91.9	91.2±0.0	**91.3±0.0**	91.9±0.0	**92.0±0.0**	mAP
		50%	56.1	63.4±0.5	**64.2±0.4**	75.9±0.2	**76.5±0.2**	CF1
			60.9	66.5±0.5	**67.2±0.5**	78.2±0.2	**78.8±0.2**	OF1
			89.2	88.2±0.0	**88.3±0.0**	90.1±0.1	**90.2±0.1**	mAP
MLP_r	pairflip	30%	80.0	80.0±0.2	**80.1±0.2**	86.5±0.1	**86.6±0.1**	CF1
			81.3	81.5±0.2	**81.6±0.2**	**86.9±0.2**	**86.9±0.3**	OF1
			87.1	86.9±0.1	**87.1±0.1**	91.1±0.1	**91.2±0.1**	mAP
		50%	50.7	54.5±0.7	**54.7±0.6**	78.9±0.7	**79.2±0.8**	CF1
			52.6	56.0±0.5	**56.2±0.5**	80.6±0.3	**80.9±0.3**	OF1
			57.7	**59.7±0.6**	**59.7±0.6**	86.1±0.3	**86.3±0.3**	mAP
	symmetric	30%	77.2	81.3±0.4	**81.8±0.3**	84.8±0.0	**85.0±0.0**	CF1
			78.4	82.0±0.2	**82.3±0.2**	85.6±0.1	**85.8±0.1**	OF1
			91.3	91.0±0.1	**91.1±0.1**	92.2±0.0	**92.3±0.0**	mAP
		50%	59.2	66.1±0.5	**67.4±0.5**	75.7±0.4	**76.4±0.3**	CF1
			60.2	66.2±0.4	**67.4±0.3**	77.6±0.3	**78.1±0.2**	OF1
			88.2	87.5±0.2	**87.7±0.2**	89.7±0.1	**89.8±0.1**	mAP
UNM_r	pairflip	30%	59.4	70.8±0.1	**71.1±0.1**	81.0±1.7	**81.2±1.6**	CF1
			58.7	68.9±0.1	**69.2±0.1**	80.8±1.8	**80.9±1.7**	OF1
			85.9	**86.0±0.1**	**86.0±0.1**	**89.4±0.1**	**89.4±0.1**	mAP
		50%	45.6	53.4±0.4	**53.6±0.5**	79.9±0.1	**80.1±0.0**	CF1
			48.4	55.6±0.8	**55.8±0.7**	80.8±1.2	**80.9±1.2**	OF1
			51.7	57.6±3.3	**57.8±3.3**	86.2±0.2	**86.3±0.2**	mAP
	symmetric	30%	78.6	81.3±0.2	**81.6±0.2**	83.7±0.2	**83.9±0.1**	CF1
			80.4	82.2±0.2	**82.4±0.2**	84.8±0.2	**84.9±0.1**	OF1
			91.2	90.4±0.0	**90.5±0.0**	91.5±0.0	**91.5±0.1**	mAP
		50%	57.8	65.8±0.5	**66.9±0.5**	76.3±0.3	**76.8±0.3**	CF1
			56.4	63.6±0.5	**64.7±0.5**	77.3±0.2	**77.7±0.2**	OF1
			87.1	87.4±0.2	**87.5±0.2**	**89.6±0.0**	**89.6±0.0**	mAP

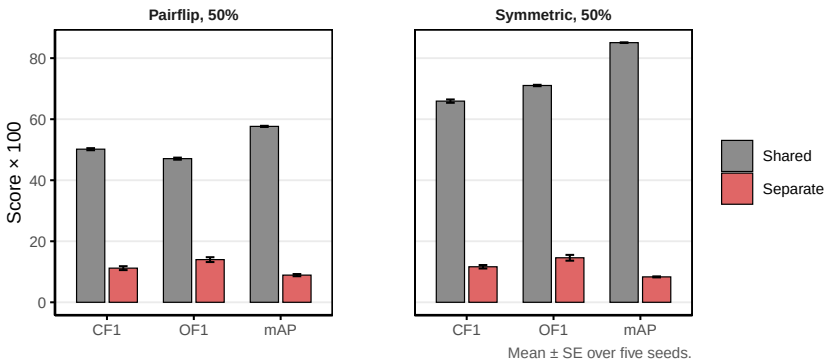


Figure 5: Shared versus separate-decoder correction on VOC07 (MLP_r, 50% noise), by noise type and metric. Bars show the mean over five repetitions and error bars mark the standard error.

tains CF1 of 48.3 (pairflip) and 67.4 (symmetric), the five-repetition runs of Table 1, with OF1 and mAP at similar or higher levels (Figure 5 plots the checkpoint-matched subset; the comparison is unaffected). The gap between the two rows is an order of magnitude for every metric and both noise types, far exceeding the repetition spread within either variant; hence the separation is not attributable to run-to-run variation. The collapse is structural rather than a tuning artifact; the unsupervised objective supervises only the noisy branch ($\hat{\mathbf{y}}$ from $\hat{\mathbf{z}}$), hence the clean decoder never receives a training signal during the unsupervised phase and cannot learn to correct. The shared decoder transfers the noisy-supervision signal to the clean branch; hence the corrected output inherits the learned clean–noisy correspondence. This directly supports the shared-manifold design.

H.6 Results on Real-World Datasets

On DeepFashion (Figure 6), LSNPC improves mAP for all six baselines (mean gain $+1.0$), and CF1 and OF1 for five of the six, with only ASL regressing slightly on the threshold metrics. The gains are largest for the weaker bases (ADDGCN, HLC), mirroring the synthetic-noise results. The improvement is smaller than on the synthetic-noise benchmarks because DeepFashion labels are comparatively clean; the method nonetheless raises the metrics without any architectural change to the base, supporting its real-world applicability.

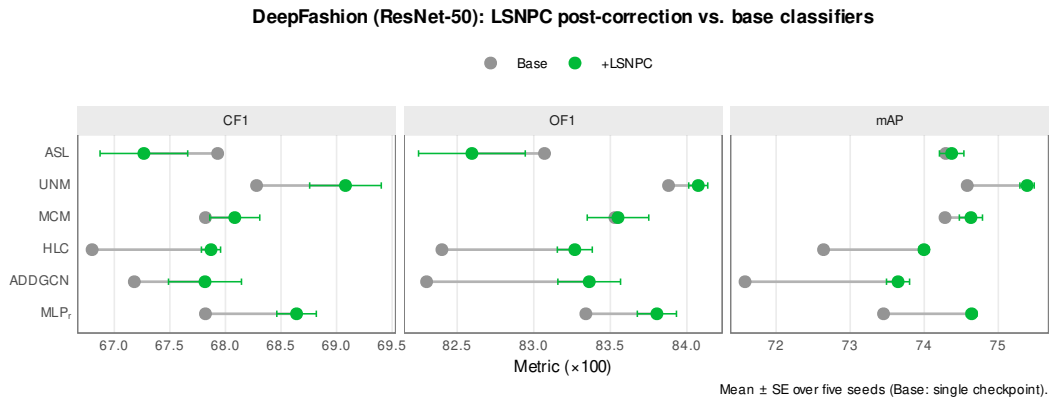


Figure 6: DeepFashion results with different baselines and LSNPC as a post-corrector. Gray: base classifier; blue: +LSNPC. Mean ± SE over five repetitions (base: single checkpoint).

I Interpreting the Learned Representation

Figure 7 visualizes the learned representations through GradCAM (Selvaraju et al., 2017), highlighting image regions that most influence predictions. Where the base classifier misses labels (blue), LSNPC shifts attention to the relevant regions, recovering signal overlooked by the noisy classifier. For misclassifications (red), it constrains focus to ground-truth regions, suppressing spurious activations. The representation thus attends to semantically meaningful areas even when the base classifier is confused, suggesting that the latent geometry encodes visual semantics rather than noise artifacts.

J Limitations

The semi-supervised variant is the most effective yet the most costly component of our framework. Even a small clean subset yields consistent further gains over the unsupervised correction, a capability none of the baselines offers; our experiments show that a modest clean fraction already suffices. Its cost is precisely the clean subset; clean annotations are scarce when labels are noisy. The marginal annotation cost is therefore the central practical tradeoff of the method, and it is well-aligned with the value it returns. When no clean data is available, the fully unsupervised variant remains effective, hence the framework degrades gracefully rather than failing. On the other hand, LSNPC's ranking behavior is not perfect. Because the ELBO optimizes pointwise reconstruction rather than within-class ordering, correction can sharpen probabilities in a way that improves the

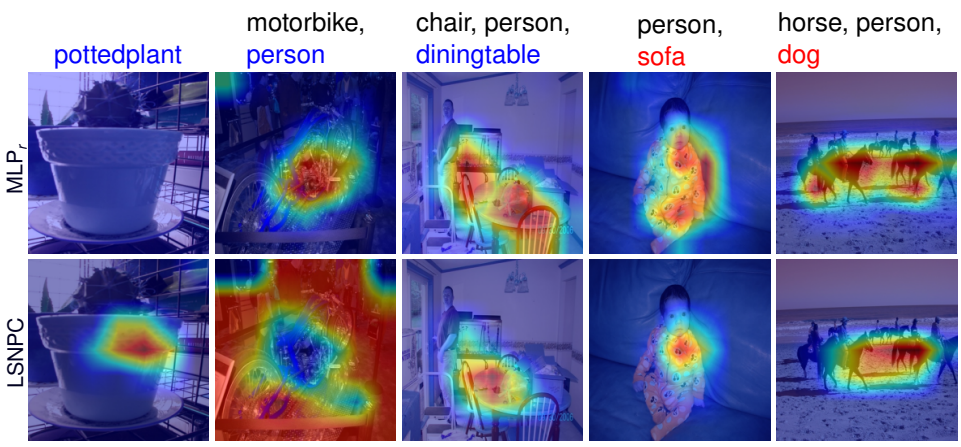


Figure 7: GradCAM visualizations comparing a base classifier (top) and LSNPC (bottom). Blue indicates labels that are recovered; red indicates labels that are corrected.

threshold-based metrics (CF1, OF1) while occasionally scrambling the ranking, degrading mAP in specific regimes: moderate symmetric noise with a strong noise-aware base classifier, whose ranking is already good. Section G characterizes precisely when this occurs and quantifies it. The gap is a property of the objective rather than a defect of the representation, and it points to a concrete improvement: a ranking-aware objective (e.g., listwise losses) would reconcile the threshold-based and ranking-based metrics while retaining reconstruction fidelity.