Pro-t-VaDE: Robust Prototype Discovery via Closed-Form Student-t Variational Deep Embedding

Zhikang Xu[1], **Jiarui Xing**[2], **Jian Wang**[3*]
[1]Institute of Intelligent Information Processing, Shanxi University
[2]School of Medicine, Yale University
[3]Boston Children's Hospital, Harvard Medical School
xuzhikangnba1@sxu.edu.cn, jiarui.xing@yale.edu, jianbljh@gmail.com

Abstract

We present a robust deep generative framework for unsupervised image prototype discovery, built on a heavy-tailed Student's t mixture prior, that bridges statistical robustness and mathematical tractability. Existing generative approaches typically rely on noise-prone Gaussian priors or computationally intensive diffusion models; neither simultaneously provides explicit latent structure nor exact analytic optimization. While diffusion models can yield strong priors, iterative sampling and intractable likelihoods limit scalability and interpretability for prototype discovery. We address these issues by constructing a Student's t latent space via a Bayesian scale-mixture formulation. This yields a fully closed-form augmented ELBO for unsupervised Student's t mixture variational autoencoders, which, to our knowledge, is the first such objective applied to robust image prototype discovery, avoiding biased variational approximations and high computational costs in prior heavy-tailed or diffusion-based frameworks. Treating precision as a latent variable automatically modulates each data point's influence, downweighting outliers to learn a data-dependent measure of trust. The result is improved robustness, accuracy, and efficiency, with well-separated latent representations and more stable manifold learning than state-of-the-art methods. We validate our model on standard vision benchmarks and challenging neuroimaging tasks. In brain MRI-driven neurodegenerative disease analysis, our model recovers clean, anatomically coherent prototypes where clinical noise often collapses rare subtypes. These results highlight the ability of our framework to preserve anatomy-associated patterns for precise discovery in clinical settings. Our code is available at: https://github.com/xuzhikangnba/torch-tvae.

1 Introduction

A core goal in unsupervised visual representation learning is to compress large image collections into a small set of *prototypes*: canonical exemplars that summarize data modes and support comparison, retrieval, and interpretation. Their utility hinges on structural integrity [33, 6]: prototypes that capture coherent modes (not averaged blur) reveal latent structure that pooled descriptors can erase. This idea appears as cluster centroids in deep clustering [33, 6, 34], object-centric representations in self-supervision [7, 8], deformable atlases in medical imaging [4, 24, 10], and morphological exemplars in computational pathology [25]. Deformable atlas building methods [24, 29] perform well when images share a consistent topology, but they frequently degrade under topological discrepancies (e.g., distinct object categories) or severe image artifacts, mainly due to their reliance on a globally

*The theoretical framework of this work was developed while the author was at Harvard Medical School; the author is currently affiliated with Eli Lilly and Company.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

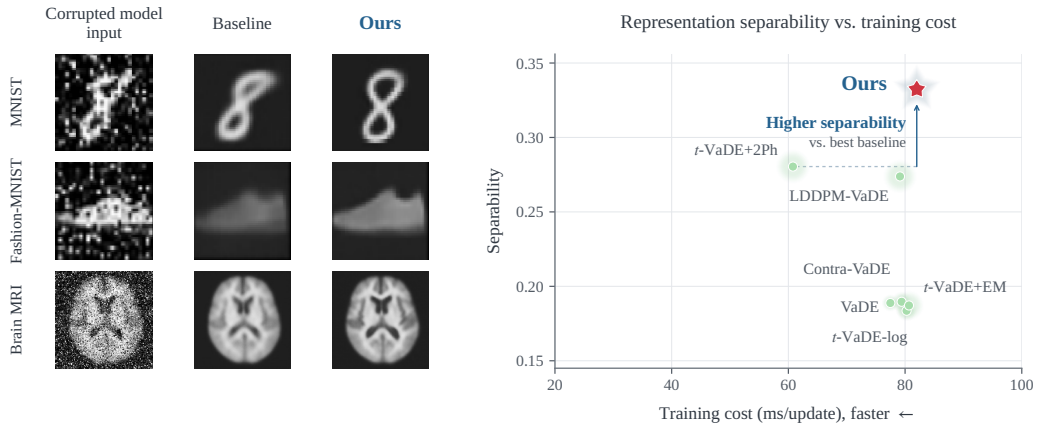


Figure 1: **Prototype generation and representation separability versus training cost.** *Left:* Noise-corrupted inputs, diffusion-based prototype estimates, and our method on MNIST, Fashion-MNIST, and brain MRI. *Right:* Mean class separability vs. per-update training time on brain scans; shading shows population variance.

defined template. In this work, we address this fundamental constraint by modeling data variation via flexible, probabilistic prototypes, defined as the component centers of a heavy-tailed generative mixture in a learned latent space, providing intrinsic robustness to topological heterogeneity and distributional perturbations without requiring an explicit template.

Two paradigms dominate deep prototype learning, largely from deep clustering. *Discriminative/contrastive* methods jointly learn representations and cluster assignments, from early deep embedded clustering to modern self-supervised frameworks [33, 8, 21, 27, 12]. However, their prototypes are usually k-means centroids, lacking an explicit probabilistic treatment of uncertainty and noise. *Generative mixture modeling*, instead, embeds an explicit latent-space mixture in a deep generative model, where each prototype is the mean of an interpretable component density [14, 15]. These methods mainly differ in objective tractability. Diffusion-based approaches either pair DDPMs with discrete clustering via EM-style updates [13, 35, 9, 30] or use diffusion as a generative regularizer [31]; in both cases, per-sample losses are typically stochastic, and clustering is often applied post hoc rather than jointly optimized. VaDE [14] is a key exception: its ELBO is fully closed form, enabling exact gradients and stable end-to-end training.

In VaDE-based frameworks, preserving analytical tractability while achieving statistical robustness is difficult. The standard Gaussian mixture prior penalizes distant latent codes quadratically, so even a few outliers can induce large gradients that pull prototypes toward spurious modes and weaken cluster separation. Robust statistics therefore motivates heavy-tailed Student's t components to downweight extremes [19, 23]. In deep generative models, however, a t prior typically removes a closed-form ELBO. Prior t-based extensions [17, 1] regain tractability via hierarchical variational bounds with hard-to-measure surrogate bias, while other heavy-tailed VAEs [15] use high-variance Monte Carlo gradients. Although an exact closed-form t-mixture was derived for supervised authorship attribution by conditioning components on class labels [5], its extension to unsupervised image prototype discovery and robustness to heavy-tailed corruption are still untested.

We close this gap by expressing the Student's t prior in its classical scale-mix form [2]: a Student's t variable is a Gaussian whose precision is modulated by a Gamma latent variable, i.e., a Gaussian-Gamma hierarchy that we previously used for subject-specific regularization in Bayesian atlas building [28]. Treating precision as an explicit per-cluster latent variable and inferring it jointly with cluster assignments yields a strict generalization of VaDE. The marginal prior over cluster means remains Student's t for heavy-tailed robustness, while conditionally it factorizes into Gaussian and Gamma terms that integrate out in closed form via conjugacy. This gives a theoretically principled, robust, and efficient generative framework for image prototype discovery and robust representation learning. As shown in Figure 1, our method substantially improves representation separability with only a modest increase in per-update training time. We summarize our core contributions as follows:

- **Theoretical Foundation.** We derive a fully closed-form training objective for an unsupervised Student's t-mixture deep generative model (a heavy-tailed mixture VAE) and, to our knowledge, are the first to apply it to image prototype discovery. We expose Gaussian–Gamma conjugacy through an auxiliary-variable construction. This formulation avoids biased variational approximations; furthermore, it ensures that the standard Gaussian-mixture case is recovered exactly as a limiting distribution when the degrees of freedom approach infinity.
- **Inherent Noise Immunity.** We show that treating precision as a latent variable naturally yields an intrinsic reweighting mechanism. Prototype gradients are automatically downweighted based on a sample's distance from its cluster center, recovering classical robust M-estimators. This enables the model to learn adaptive, per-cluster degrees of freedom that handle varying noise levels without manual hyperparameter tuning.
- **Empirical Performance.**The proposed model achieves higher prototype assignment accuracy and more distinctly separated latent representations, improves training stability over state-of-the-art baselines on standard image benchmarks, matches the training cost of Gaussian VaDE or single-step generative models, and enables 33–36$\times$ faster inference than a latent-diffusion baseline (e.g., with a step size exceeding 100). Learned prototypes preserve class-specific features under additive and structured noise, while baseline generative methods often collapse into blurred, non-discriminative templates.
- **Clinical Utility.** In real-world clinical practice (e.g., during the diagnostic workup of neurodegenerative disorders), the model recovers anatomically faithful prototypes and well-calibrated uncertainty in cluster assignments. By preserving rare subtypes in high-noise regimes, it supports precise clinical analysis of disease heterogeneity where standard models might otherwise merge distinct atrophy patterns.

2 Background: Variational Deep Embedding

Mathematical Notation. We consider the task of embedding observations $\boldsymbol{x} \in \mathbb{R}^D$ into a latent space $\boldsymbol{z} \in \mathbb{R}^d$ characterized by mixture-component (prototype) labels $c \in \{1, \dots, K\}$. We denote a Gaussian distribution as $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1})$ with mean $\boldsymbol{\mu}$ and precision $\boldsymbol{\Lambda}$. The Student distribution $\mathcal{T}_\nu(\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1})$ is defined by degrees of freedom ν, location $\boldsymbol{\mu}$, and precision $\boldsymbol{\Lambda}$. We adopt the parameterization of the shape-rate for the Gamma distribution, where $u \sim \mathrm{Gamma}(\alpha, \beta)$ denotes the density $p(u) \propto u^{\alpha-1} e^{-\beta u}$ with the mean α/β. The digamma function is denoted by ψ.

The VaDE Framework. Variational Deep Embedding (VaDE) [14] defines a generative process where cluster assignments and latent codes are sampled as follows:

$$c \sim \mathrm{Cat}(\boldsymbol{\pi}), \quad \boldsymbol{z} \mid c \sim \mathcal{N}(\boldsymbol{\mu}_c, \boldsymbol{\Lambda}_c^{-1}), \quad \boldsymbol{x} \mid \boldsymbol{z} \sim p_\theta(\boldsymbol{x} \mid \boldsymbol{z}). \tag{1}$$

Here, $\mathrm{Cat}(\boldsymbol{\pi})$ is the categorical distribution over K components, with mixing probabilities $\pi_c \geq 0$ and $\sum_{c=1}^K \pi_c = 1$. The inference model is parameterized by a diagonal Gaussian encoder $q_\phi(\boldsymbol{z} \mid \boldsymbol{x}) = \mathcal{N}(\boldsymbol{\mu}_\phi, \mathrm{diag}(\boldsymbol{\sigma}_\phi^2))$. Cluster responsibilities $\gamma_c := q(c \mid \boldsymbol{x})$ are computed via Bayes' rule

$$\gamma_c = \frac{\pi_c \, p(\boldsymbol{z} \mid c)}{\sum_{k=1}^K \pi_k \, p(\boldsymbol{z} \mid k)}, \tag{2}$$

where $\boldsymbol{z}$ is a sample from the encoder. The lower bound of the VaDE evidence (ELBO) is given by:

$$\mathcal{L}_{\text{VaDE}}(\boldsymbol{x}) = \mathbb{E}_{q_\phi}[\log p_\theta(\boldsymbol{x} \mid \boldsymbol{z})] - \sum_{c=1}^K \gamma_c \left[\mathrm{KL}[q_\phi(\boldsymbol{z} \mid \boldsymbol{x}) \,\|\, \mathcal{N}(\boldsymbol{\mu}_c, \boldsymbol{\Lambda}_c^{-1})] - \log \pi_c \right]. \tag{3}$$

This objective is computationally efficient because the KL divergence between two Gaussians admits a closed-form expression [11, 16].

The Analytic Obstacle for Robust Priors. Replacing the Gaussian prior in (1) with a Student's t prior $\mathcal{T}_{\nu_c}(\boldsymbol{\mu}_c, \boldsymbol{\Lambda}_c^{-1})$ introduces robustness to outliers due to the distribution's heavy tails [22]. However, this substitution renders the KL divergence in (3) intractable, as the integral $\int q_\phi(\boldsymbol{z} \mid \boldsymbol{x}) \log \mathcal{T}_{\nu_c}(\boldsymbol{z}) d\boldsymbol{z}$ has no closed-form solution. Existing methods often rely on numerical approximations or moment-matching surrogates [17, 15], which may introduce gradient variance or bias; a notable exception is the supervised text model of [5]. We present a methodology that retains a strictly closed-form ELBO under Student's t priors.

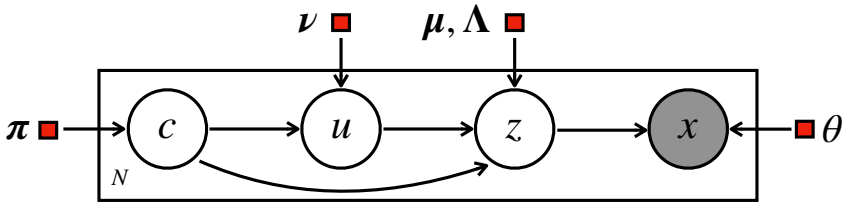


Figure 2: Probabilistic graphical model of **Pro-t-VaDE** as a *scale mixture of Gaussians*. Shaded nodes are observed, white nodes latent, red squares global parameters, and plate N indicates per-sample replication. Given cluster c, auxiliary precision $u \sim \mathrm{Gamma}(\nu_c/2, \nu_c/2)$ scales $\boldsymbol{z}$ covariance to $(u\boldsymbol{\Lambda}_c)^{-1}$: small u models tails, large u concentrates near the center. Marginalizing u yields the Student's t prior $\boldsymbol{z} \mid c \sim \mathcal{T}_{\nu_c}(\boldsymbol{\mu}_c, \boldsymbol{\Lambda}_c^{-1})$. The encoder maps $\boldsymbol{x}$ to $q_\phi(\boldsymbol{z} \mid \boldsymbol{x})$, the mixture gives soft prototype assignments, and the shared decoder maps $\boldsymbol{z}$ to reconstructions and $\boldsymbol{\mu}_c$ to image prototypes.

3 Methodology

3.1 Generative Model

Our construction rests on a classical scale-mixture identity [2], central to robust variational Bayes (VB) mixtures [26, 23] and hierarchical regularization priors in medical image analysis [28]:

$$u \sim \mathrm{Gamma}\big(\tfrac{\nu}{2}, \tfrac{\nu}{2}\big), \quad \boldsymbol{z} \mid u \sim \mathcal{N}\big(\boldsymbol{\mu}, (u\boldsymbol{\Lambda})^{-1}\big) \quad \Longrightarrow \quad \boldsymbol{z} \sim \mathcal{T}_\nu\big(\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1}\big). \tag{4}$$

The auxiliary precision u plays a dual role. Analytically, it decomposes the heavy-tailed prior into a Gaussian–Gamma hierarchy that admits closed-form inference. Statistically, it acts as a per-sample inlier weight: large u tightens the conditional Gaussian around an inlier observation, while small u inflates it to absorb tail observations without biasing the cluster mean. We adopt this identity in a per-cluster form to obtain the hierarchical generative model

$$c \sim \mathrm{Cat}(\boldsymbol{\pi}), \quad u \mid c \sim \mathrm{Gamma}\big(\tfrac{\nu_c}{2}, \tfrac{\nu_c}{2}\big), \quad \boldsymbol{z} \mid u, c \sim \mathcal{N}\big(\boldsymbol{\mu}_c, (u\boldsymbol{\Lambda}_c)^{-1}\big), \quad \boldsymbol{x} \mid \boldsymbol{z} \sim p_\theta(\boldsymbol{x} \mid \boldsymbol{z}), \tag{5}$$

illustrated in Fig. 2. Marginalizing u recovers the per-cluster Student's t prior $\boldsymbol{z} \mid c \sim \mathcal{T}_{\nu_c}(\boldsymbol{\mu}_c, \boldsymbol{\Lambda}_c^{-1})$, so (5) is statistically equivalent to a t-mixture VaDE while preserving conditionally Gaussian latents. The degrees of freedom $\nu_c > 0$ are learned per cluster, so each prototype adapts its tail behavior to the empirical noise level of its cluster.

3.2 Closed-Form ELBO via Analytical Collapsing

Substituting the Student-t prior of (5) into the variational objective introduces a fundamental obstacle: the marginal divergence $\mathrm{KL}\big[q_\phi(\boldsymbol{z} \mid \boldsymbol{x}) \,\|\, \mathcal{T}_{\nu_c}(\boldsymbol{\mu}_c, \boldsymbol{\Lambda}_c^{-1})\big]$ admits no closed form since it requires evaluating $\mathbb{E}_{q_\phi}\big[\log(1 + \bar{\Delta}_c^2/\nu_c)\big]$. We resolve this *variational intractability* through the *analytical collapsing* of the auxiliary variable u, exploiting Gaussian–Gamma conjugacy to recover a fully analytic ELBO with exact gradients.

We employ the structured variational posterior

$$q(\boldsymbol{z}, u, c \mid \boldsymbol{x}) \;=\; q_\phi(\boldsymbol{z} \mid \boldsymbol{x})\, q(u \mid \boldsymbol{x}, c)\, q(c \mid \boldsymbol{x}), \tag{6}$$

in which $q_\phi(\boldsymbol{z} \mid \boldsymbol{x}) = \mathcal{N}\big(\boldsymbol{\mu}_\phi(\boldsymbol{x}), \mathrm{diag}(\boldsymbol{\sigma}_\phi^2(\boldsymbol{x}))\big)$ is the amortized encoder and $\gamma_c \equiv q(c \mid \boldsymbol{x})$ is the cluster posterior. Defining the expected squared Mahalanobis distance under q_ϕ,

$$\bar{\Delta}_c^2 \;\equiv\; \sum_{j=1}^{d} \Lambda_{c,j} \left[\sigma_{\phi,j}^2 + (\mu_{\phi,j} - \mu_{c,j})^2\right], \tag{7}$$

Gaussian–Gamma conjugacy yields the optimal posterior in closed form,

$$q^\star(u \mid \boldsymbol{x}, c) = \mathrm{Gamma}(\tilde{\alpha}_c, \tilde{\beta}_c), \qquad \tilde{\alpha}_c = \tfrac{\nu_c + d}{2}, \qquad \tilde{\beta}_c = \tfrac{\nu_c + \bar{\Delta}_c^2}{2}, \tag{8}$$

with sufficient statistics

$$\mathbb{E}[u] \;=\; \frac{\nu_c + d}{\nu_c + \bar{\Delta}_c^2}, \qquad \mathbb{E}[\log u] \;=\; \psi(\tilde{\alpha}_c) - \log \tilde{\beta}_c. \tag{9}$$

Substituting $q^\star$ into the joint variational objective and integrating u analytically, we obtain the central result of this section: a fully closed-form ELBO for the Student's t-mixture model,

$$\mathcal{L}(\boldsymbol{x}) \;=\; \mathbb{E}_{q_\phi}\big[\log p_\theta(\boldsymbol{x} \mid \boldsymbol{z})\big] \;-\; \sum_{c=1}^{K} \gamma_c \left[\widetilde{\mathrm{KL}}^{(c)}_{\mathrm{Gauss}} + \mathrm{KL}^{(c)}_{\mathrm{Gamma}}\right] \;-\; \mathrm{KL}\big[q(c \mid \boldsymbol{x}) \,\|\, p(c)\big], \tag{10}$$

All terms are evaluated analytically, with no sampling of u or KL approximations. Because the KL for the marginal Student-t distribution $\boldsymbol{z} \mid c$ is intractable, it must be split into two KL terms. Introducing an auxiliary variable yields a tractable pair $(\widetilde{\mathrm{KL}}_{\mathrm{Gauss}}, \mathrm{KL}_{\mathrm{Gamma}})$ whose expected sum upper-bounds the marginal KL and provides exact gradients. The three terms of (10) are

$$\widetilde{\mathrm{KL}}^{(c)}_{\mathrm{Gauss}} = \tfrac{1}{2}\Big[\mathbb{E}[u]\,\bar{\Delta}_c^2 \;-\; d\,\mathbb{E}[\log u] \;-\; \textstyle\sum_j \log \Lambda_{c,j} \;-\; \sum_j \log \sigma^2_{\phi,j} \;-\; d\Big], \tag{11}$$

$$\mathrm{KL}^{(c)}_{\mathrm{Gamma}} = \tfrac{d}{2}\psi(\tilde{\alpha}_c) \;+\; \log \tfrac{\Gamma(\nu_c/2)}{\Gamma(\tilde{\alpha}_c)} \;+\; \tfrac{\nu_c}{2}\log\Big(1 + \tfrac{\bar{\Delta}_c^2}{\nu_c}\Big) \;-\; \tfrac{1}{2}\mathbb{E}[u]\,\bar{\Delta}_c^2, \tag{12}$$

$$\mathrm{KL}\big[q(c \mid \boldsymbol{x}) \,\|\, p(c)\big] = \textstyle\sum_{c=1}^{K} \gamma_c\big(\log \gamma_c - \log \pi_c\big). \tag{13}$$

Eq. (11) is a Gaussian KL where the prior precision is scaled by the inferred inlier weight $\mathbb{E}[u]$: distant samples have smaller $\mathbb{E}[u]$ and thus a weaker influence on the cluster mean, matching the Iteratively Reweighted Least Squares (IRLS) dynamics in (14). Eq. (12) adjusts the per-cluster degrees of freedom via $\frac{\nu_c}{2}\log\big(1 + \bar{\Delta}_c^2/\nu_c\big)$, the Student's t tail penalty, so ν_c can be learned by gradient descent instead of being fixed. Eq. (13) is the standard categorical divergence between the cluster posterior and prior. Full derivations, expectation identities, and the extension to full-covariance $\boldsymbol{\Lambda}_c$ appear in the appendix.

Computational Cost. Notably, The auxiliary scale variable needs no extra networks or Monte Carlo estimates, since $\mathbb{E}[u]$ and $\mathbb{E}[\log u]$ have closed forms. Compared with Gaussian VaDE, it adds only K trainable scalars. Overhead scales as $O(BK)$, while diagonal mixture evaluation remains $O(BKd)$ for batch size B, K components, and latent dimension d. Thus, it improves robustness to heavy-tailed noise with essentially the same asymptotic complexity as a Gaussian mixture prior.

Intrinsic Robustness and Gradient Dynamics. The model's robustness follows from the gradient of (10) w.r.t. $\boldsymbol{\mu}_c$:

$$\nabla_{\boldsymbol{\mu}_c}\mathcal{L} \propto \gamma_c\,\mathbb{E}[u]\,\boldsymbol{\Lambda}_c(\boldsymbol{\mu}_\phi - \boldsymbol{\mu}_c), \quad \text{where} \quad \mathbb{E}[u] = \frac{\nu_c + d}{\nu_c + \bar{\Delta}_c^2}. \tag{14}$$

This gives an IRLS scheme [19]. For inliers with $\bar{\Delta}_c^2 \approx d$, $\mathbb{E}[u] \approx 1$, recovering the Gaussian VaDE update. For outliers with $\bar{\Delta}_c^2 \gg \nu_c$, $\mathbb{E}[u] \to 0$, so they have negligible effect on the cluster prototype.[2] Unlike standard IRLS, these weights come from the variational objective, not an imposed rule.

Plug-and-Play Generality. The conjugate construction can be incorporated into encoder-decoder architectures satisfying the stated latent-distribution assumptions. Our experiments use ResNet-based backbones, and the evaluation of method with additional architectures is future work. A high-level forward pass is given in Alg. 1. In this sense, the Student's t mixture prior acts as a drop-in, closed-form replacement for the Gaussian mixture prior of VaDE-style generative models.

[2]This matches classical IRLS for robust t-mixture estimation, where $w_i \propto (\nu + d)/(\nu + \Delta_i^2)$ and prototypes are updated by IRLS [19, 23, 3], yielding vanishing outlier influence.

Algorithm 1 Forward pass of our model.

Require: Input $\boldsymbol{x}$; encoder q_ϕ; decoder p_θ; mixture parameters $\{\boldsymbol{\mu}_c, \boldsymbol{\Lambda}_c, \pi_c, \nu_c\}_{c=1}^{K}$.

$(\boldsymbol{\mu}_\phi, \boldsymbol{\sigma}_\phi^2) \leftarrow \mathrm{Enc}(\boldsymbol{x})$; $\quad \boldsymbol{z} \leftarrow \boldsymbol{\mu}_\phi + \boldsymbol{\sigma}_\phi \odot \boldsymbol{\epsilon}, \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$.

$\hat{\boldsymbol{x}} \leftarrow \mathrm{Dec}(\boldsymbol{z})$; $\quad \mathcal{L}_{\mathrm{rec}} \leftarrow -\log p_\theta(\boldsymbol{x} \mid \boldsymbol{z})$.

$\gamma_c \propto \pi_c \, \mathcal{T}_{\nu_c}(\boldsymbol{z} \mid \boldsymbol{\mu}_c, \boldsymbol{\Lambda}_c^{-1})$ for all c.

for $c = 1, \ldots, K$ **do**

  Compute $\bar{\Delta}_c^2$ via (7); $\mathbb{E}[u], \mathbb{E}[\log u]$ via (9).

  Evaluate $\widetilde{\mathrm{KL}}_{\mathrm{Gauss}}^{(c)}$ (11) and $\mathrm{KL}_{\mathrm{Gamma}}^{(c)}$ (12).

end for

Compute $\mathrm{KL}_{\mathrm{Cat}} \leftarrow \sum_c \gamma_c \big(\log \gamma_c - \log \pi_c\big)$ via (13).

return $\mathcal{L} = \mathcal{L}_{\mathrm{rec}} + \sum_c \gamma_c \left[\widetilde{\mathrm{KL}}_{\mathrm{Gauss}}^{(c)} + \mathrm{KL}_{\mathrm{Gamma}}^{(c)}\right] + \mathrm{KL}_{\mathrm{Cat}}$.

4 Experiments

We evaluate our method on three axes: prototype assignment accuracy under noise (§4.2), prototype quality and representation separability (§4.3), and clinical validity in a real-world neurodegenerative cohort (§4.4). We describe the setup, datasets, baselines, and evaluation protocols once below; later sections report only results.

4.1 Setup

We evaluate on two standard image benchmarks and one real-world medical imaging cohort. MNIST [20] and Fashion-MNIST [32] both contain $60{,}000$ training and $10{,}000$ test grayscale images of size 28×28 across $K = 10$ classes; together they probe prototype discovery data of qualitatively different complexity (handwritten digits vs. clothing categories). OASIS Brain MRI [18] provides T_1-weighted brain volumes of size 128^3 from a subset of 253 subjects aged 60–95, including healthy controls and Alzheimer's disease patients, used to assess the model's ability to recover anatomically faithful disease prototypes under realistic acquisition noise.

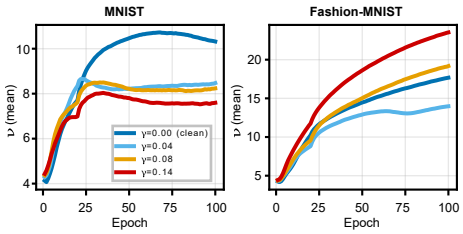
Baselines. We benchmark against six representative deep clustering and generative mixture models spanning key architectural and methodological choices. VaDE [14] is a Gaussian-prior variational deep clustering model. Contra-VaDE [21] adds a contrastive objective to separate latent clusters. LDDPM-VaDE [35] adapts ClusterDDPM by running diffusion in latent space. The other three baselines use a Student's t prior, addressing variational intractability differently: t-VaDE-log uses Monte Carlo log-gradient estimators for $\mathbb{E}_{q_\phi}[\log(1 + \bar{\Delta}^2/\nu)]$ [15]; t-VaDE-EM alternates variational updates with EM cluster-parameter updates [17]; and t-VaDE-2Ph trains in two phases, transitioning from a Gaussian to a Student's t prior. Our method is the only fully end-to-end approach with a closed-form ELBO under a Student's t prior.

Implementation. All baselines and our model use the same ResNet-based encoder–decoder backbone with a symmetric decoder. We train for 100 epochs using AdamW with a learning rate of 10^{-3} for all parameters, weight decay 10^{-4}, and batch size 32. The number of prototypes is fixed to the ground-truth $K = 10$ on MNIST and Fashion-MNIST and $K = 2$ for the brain MRI dataset. In our model, ν_c is softplus-parameterized and initialized at $\nu_c = 5$ for every component. Sensitivity to the initialization of ν_c can be found in the appendix. All experiments run on $4 \times$ NVIDIA 5880 Ada GPUs with 48 GB memory each.

Evaluation protocols. All experiments follow the same protocol. For *prototype assignment accuracy* (§4.2), we corrupt MNIST and Fashion-MNIST with Cauchy noise at scales $\gamma \in \{0.02, 0.04, \ldots, 0.14\}$ and report standard clustering metrics ACC, ARI, and NMI (higher is better). Cauchy noise is a strict stress test because its undefined moments violate Gaussian-prior assumptions of finite second-order statistics. For *prototype quality* (§4.3), under the same noise we measure: (i) PSNR between matched prototypes learned from noisy vs. clean data; (ii) latent separability using Silhouette (Sil, ↑), Davies–Bouldin (DB, ↓), and t-SNE; and (iii) training stability via the rate of *cluster collapse* (two or more prototypes converging to the same latent location) over 118 runs.

For *clinical validity* (§4.4), we apply nine realistic Brain MRI corruptions (Gaussian, Rician, and bias-field; mild/medium/severe) and evaluate the same metrics.

4.2 Prototype Assignment Accuracy Under Noise



Algorithm	MNIST ($K=10$) ACC↑	MNIST ($K=10$) ARI↑	MNIST ($K=10$) NMI↑	Fashion-MNIST ($K=10$) ACC↑	Fashion-MNIST ($K=10$) ARI↑	Fashion-MNIST ($K=10$) NMI↑
VaDE	0.638	0.408	0.582	0.387	0.214	0.373
Contra-VaDE	0.643	0.410	0.593	0.390	0.212	0.361
LDDPM-VaDE	0.289	0.114	0.250	0.178	0.020	0.122
t-VaDE-log	0.727	0.575	0.656	0.385	0.206	0.364
t-VaDE+EM	0.651	0.494	0.618	0.382	0.205	0.333
t-VaDE+2Ph	0.661	0.476	0.610	0.376	0.202	0.348
Ours	**0.798**	**0.695**	**0.753**	**0.555**	**0.409**	**0.570**

Figure 3: **Left:** Mean of the learned per-cluster degrees of freedom ν_c across training epochs on MNIST and Fashion-MNIST under four Cauchy-noise levels $\gamma \in \{0.00, 0.04, 0.08, 0.14\}$; all runs initialize $\nu_c = 5$. **Right:** Prototype assignment accuracy (ACC, ARI, NMI) averaged over noise levels $\gamma \in \{0.02, \ldots, 0.14\}$ on MNIST and Fashion-MNIST. Best in **bold**, second underlined.

Fig. 3 (left) tracks ν_c during training. On MNIST, ν_c converges near 10 on clean data and decreases to about 4 at $\gamma = 0.14$, consistent with robust t-estimation [19, 23]: heavier tails call for smaller ν_c to absorb outliers without biasing the prototypes. Fashion-MNIST shows the opposite trend, with ν_c rising as noise grows. This reflects its cluster geometry: classes like *pullover*, *shirt*, and *coat* already exhibit large intra-class variation, so ν_c stays small even on clean inputs; adding noise triggers the IRLS reweighting in (14) to downweight corrupted samples, tightening the remaining geometry and pushing ν_c toward the Gaussian regime. These results show ν_c adapts to dataset geometry, a capability absent in fixed-ν formulations.

Fig. 3 (right) reports prototype assignment accuracy averaged over Cauchy-noise levels $\gamma \in \{0.02, \ldots, 0.14\}$. Our method leads on every metric across both datasets. Variants with fixed or externally scheduled ν (t-VaDE-log, t-VaDE+EM, t-VaDE+2Ph) improve over VaDE on MNIST but stall on the more heterogeneous Fashion-MNIST, while LDDPM-VaDE collapses under noise on both. The gap is largest on Fashion-MNIST, where per-cluster adaptation of ν_c delivers a $\sim$42% relative ACC gain over the strongest baseline, confirming that jointly learning ν_c, rather than tuning a global heaviness parameter, is what drives robustness across heterogeneous geometries. Over five random seeds, our method achieves statistically significant gains over the strongest baselines: +0.079 on MNIST (95% CI [0.052, 0.106], p=0.0025) and +0.116 on Fashion-MNIST (95% CI [0.032, 0.200], p=0.018). See Appendix for the five-seed results and paired tests with Holm correction.

Fig. 4 (top) reports ACC, ARI, and NMI as a function of noise level on both benchmarks; across all noise scales, our method (red) maintains a near-flat performance curve, while every baseline degrades sharply and most fail for $\gamma \geq 0.08$. Averaged over the seven noise levels, our model attains ACC of 0.798 on MNIST and 0.555 on Fashion-MNIST against best-baseline values of 0.727 and 0.390 respectively, with similarly large margins on ARI and NMI.

Runtime Efficiency. The auxiliary-variable construction has a modest measured cost: training takes 33.21/32.00 ms per update on MNIST/Fashion-MNIST, compared with 27.74/27.05 for VaDE and 31.99/30.20 for LDDPM-VaDE. In a separate benchmark executing the complete 100-step reverse trajectory of the same latent-diffusion network, end-to-end assignment takes 0.142/0.142 ms per sample for LDDPM-VaDE and 0.004/0.004 for ours, corresponding to 33×/36× speedups. These speedups concern the specified reverse-inference path, not training updates. Protocols and training costs are reported in the appendix. Together with the accuracy gains above, this shows that our generative framework attains heavy-tailed robustness without sacrificing efficiency.

4.3 Prototype Quality and Representation Separability

Fig. 4 (bottom) shows decoded prototypes at $\gamma = 0.12$. Our method preserves recognizable digits on MNIST and distinguishable garments on Fashion-MNIST, while baseline prototypes are blurred or exhibit cluster collapse, with multiple prototypes converging to the same shape and leaving some classes unrepresented.

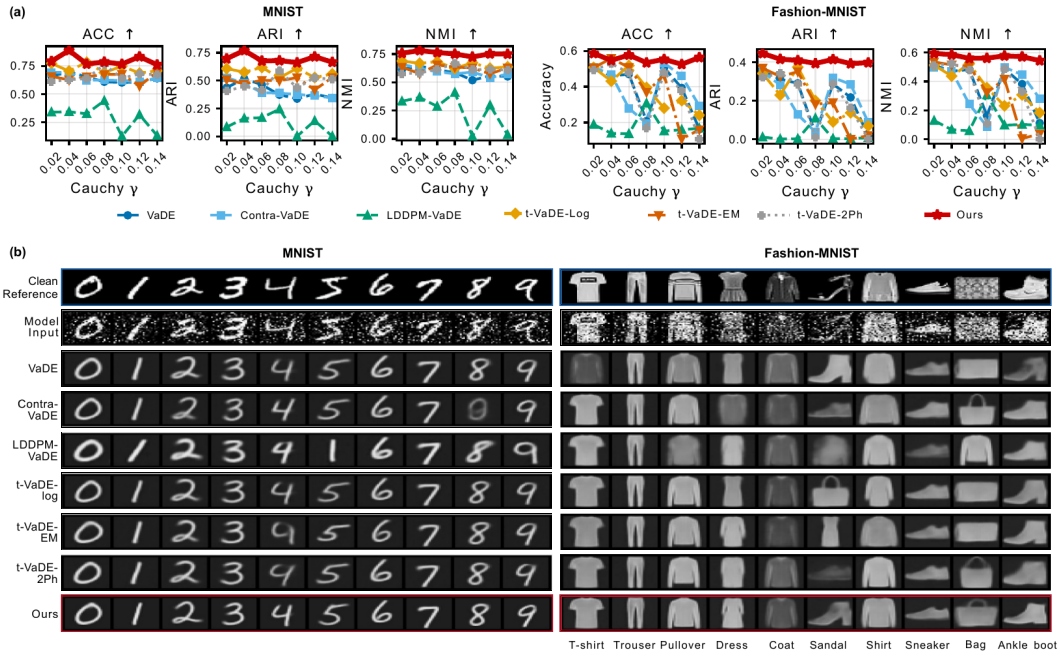


Figure 4: Top: quantitative evaluation of all models under increasing noise corruption of images. Bottom: prototype visualization across all models when image is under severe corruption $\gamma = 0.12$. Our approach preserves crisp, distinguishable prototypes without duplication.

Fig. 5 (top) shows t-SNE projections. On clean inputs, competing models generally recover cluster structure. However, under noise, VaDE leaves clusters overlapping, LDDPM-VaDE collapses into an unstructured blob, and t-VaDE-log retains only blurred boundaries, while our model preserves well-separated clusters. The per-cluster ν_c downweights corrupted samples, anchoring both prototypes and latent geometry on clean class structure rather than noise. The separated latent representations are thus a direct by-product of the heavy-tailed generative prior and underpin the quality of the discovered prototypes.

4.4 Anatomical Robustness and Medical Image Validation

Fig. 6 shows prototypes learned from heavily noise-corrupted OASIS Brain MRI (left). Our prototypes are clean and artifact-free, with sharp sagittal and coronal boundaries. VaDE, Contra-VaDE, and LDDPM-VaDE yield blurred prototypes with smeared cortex and washed-out ventricles; t-VaDE variants recover more anatomy but retain noise and cortical ghosting. Downweighting corrupted voxels via per-cluster ν_c anchors prototypes to clean anatomy rather than noise patterns, which may benefit downstream tasks such as atlas-based registration and population-level subtype analysis.

Fig. 5 (bottom-left) shows Silhouette scores on OASIS Brain MRI across nine realistic corruptions (Gaussian, Rician, and bias-field noise) at mild, medium, and severe levels. Our method attains the highest Silhouette in all conditions, with the largest gains under severe noise (G-sev, R-sev, Rg-sev). Robustness on synthetic Cauchy noise thus extends to clinically realistic artifacts, where labels are unavailable and latent separability is the main evaluation criterion.

Fig. 5 (bottom-right) shows per-pixel standard deviation of four groups of anatomy-associated patterns, each produced by averaging 32 decoded samples. From left to right, the estimated variance grows in atrophied regions such as the ventricles and surrounding cortical tissue. The heavy-tailed t-distribution thus yields not only robustness but also may potentially benefit clinically interpretable epistemic uncertainty, flagging regions where pathological change reduces the model's confidence.

Tab. 1(a) reports prototype separability and reconstruction quality on OASIS at $K = 2$, with metrics marginalized over the principal anatomical orientations to ensure the representativeness of the global volumetric structure. Our method consistently attains the best Sil and DB across all noise regimes, with both metrics remaining stable under increasing corruption γ (Sil: 0.321–0.351; DB: 1.099–

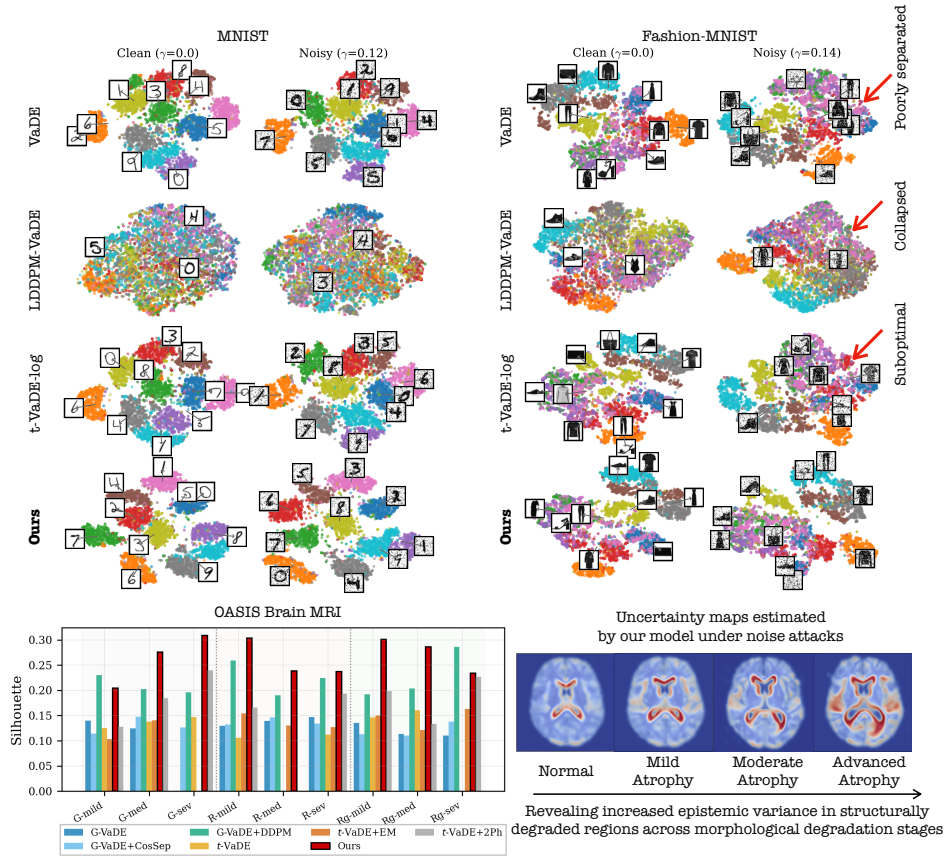


Figure 5: Top: t-SNE of latent representations from VaDE, LDDPM-VaDE, t-VaDE-log, and ours on MNIST and Fashion-MNIST with clean ($\gamma = 0.0$) and noisy inputs ($\gamma = 0.12$ MNIST, 0.14 Fashion-MNIST). Bottom-left: Silhouette on OASIS Brain MRI under nine realistic corruptions (Gaussian, Rician, bias-field; mild/medium/severe). Bottom-right: per-voxel uncertainty maps from our model across four groups of anatomy-associated patterns.

Table 1: Prototype separability and reconstruction quality on OASIS at $K = 2$, averaged over sagittal/axial/coronal views. **(a)** Sil, DB, and PSNR at four noise levels; PSNR is reported only for corrupted conditions relative to matched clean-run prototypes (clean shown as -). Collapsed runs (single cluster; Sil undefined) are excluded; cells with any excluded run are *italic*†. **(b)** Collapse count and rate per method over 118 runs. Best in **bold**.

(a) Prototype separability and reconstruction at four noise levels.

(b) Model collapse over 118 runs.

Method	$\gamma = 0$ (clean)			$\gamma = 0.02$			$\gamma = 0.08$			$\gamma = 0.12$			Count↓	Rate (%)↓
	Sil↑	DB↓	PSNR↑	Sil↑	DB↓	PSNR↑	Sil↑	DB↓	PSNR↑	Sil↑	DB↓	PSNR↑		
VaDE	0.170	1.913	–	0.165	1.952	25.75	0.181	1.934	21.81	0.200	1.811	19.82	2	1.7
Contra-VaDE	0.172	1.932	–	0.168	1.966	22.02	0.176	2.039	20.45	0.215	1.759	19.57	1	0.8
LDDPM-VaDE	0.238	1.618	–	0.249	1.534	**28.28**	0.262	1.446	21.71	*0.309*†	*1.300*†	*19.74*†	16	13.6
t-VaDE-log	0.165	1.979	–	0.164	2.013	25.38	0.183	1.965	21.88	0.176	1.923	19.47	2	1.7
t-VaDE+EM	0.167	2.055	–	0.182	1.916	24.80	0.176	1.955	21.32	0.206	1.874	19.61	4	3.4
t-VaDE+2Ph	0.248	1.520	–	0.258	1.493	25.47	0.267	1.443	**22.12**	0.242	1.608	18.93	3	2.5
Ours	**0.326**	**1.181**	–	**0.321**	**1.157**	26.15	**0.328**	**1.185**	22.05	**0.351**	**1.099**	**20.01**	**0**	**0.0**

1.185), while baselines degrade substantially. On reconstruction fidelity, our approach yields the highest PSNR (20.01 dB) at ($\gamma = 0.12$). The elevated Sil and DB of LDDPM-VaDE at $\gamma = 0.12$ are artifacts of selective aggregation (marked †): that baseline frequently suffered mode collapse, and the reported values are computed only over its surviving runs.

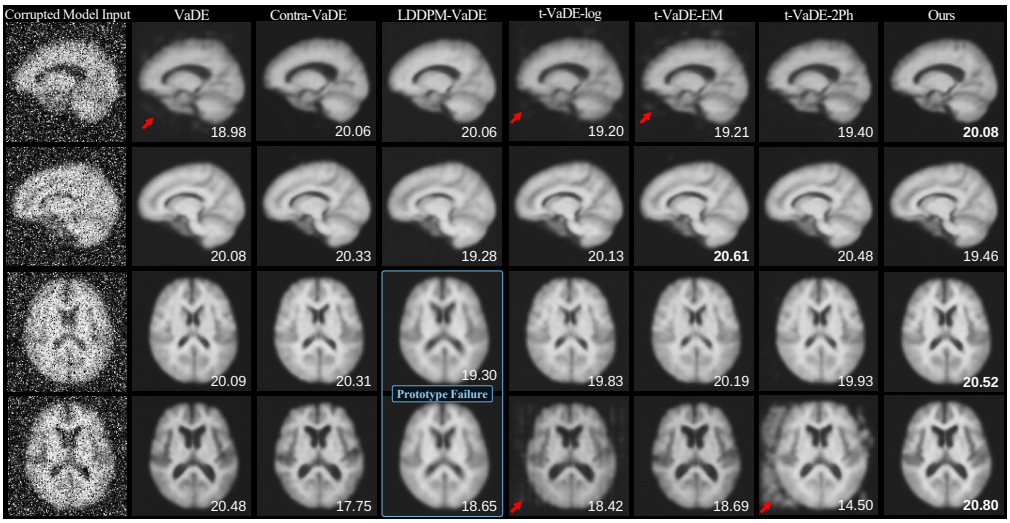


Figure 6: Visual comparison of learned atlases on OASIS dataset under Cauchy noise ($\gamma = 0.12$). Col. 1 shows corrupted inputs; cols. 2–8 show cluster-mean atlases reconstructed by different methods. Rows 1–2 are sagittal atlases for two clusters; rows 3–4 are coronal atlases. White numbers (bottom right) report per-cluster PSNR (dB) vs. each method's clean atlas after one-to-one cluster matching (higher is better); best per row is bold. Red arrows mark baseline artifacts. The blue box shows an LDDPM-VaDE failure: the two coronal cluster atlases are nearly identical, indicating prototype collapse. Aggregate PSNR and Silhouette results are reported in Tab. 1(a).

Tab. 1(b) reports cluster-collapse rates across all 118 trials. LDDPM-VaDE collapses on 13.6% of runs, the remaining baselines on 0.8%–3.4%, and our method on none. Combined with the stable Sil and DB above, this indicates that the per-cluster ν_c provides a reliability margin invisible to accuracy metrics alone. In a clinical setting, where each subject is a real patient and a degenerate solution must be discarded rather than retried, a zero-collapse rate is the property that matters most.

5 Discussion & Conclusion

In this paper, we introduce Pro-t-VaDE and show that heavy-tailed robustness and analytical tractability can be jointly achieved in deep generative mixture models for image prototype discovery.By leveraging the Gaussian-Gamma conjugate structure of the Student's t-distribution, we derive a fully closed-form ELBO, which effectively eliminates the stochastic variance inherent in traditional gradient-based variational inference. Our framework establishes a rigorous link between deep latent variable models and classical M-estimation, providing a principled mechanism to navigate the trade-off between model expressivity and robustness. Across natural-image benchmarks and brain MRI, this yields accurate, efficient, and well-separated representations with a negligible overhead over a Gaussian mixture prior. Empirical evaluations on neurodegenerative disease datasets show that our method preserves fine-grained anatomy-associated patterns that are often obscured by Gaussian-based models, offering a more granular lens into disease heterogeneity. It is worth noting that several aspects warrant future exploration. Our method is particularly suited for scenarios where data contain heavy-tailed noise yet analytical tractability is desired, such as clinical MRI with acquisition artifacts, where the per-cluster ν_c automatically downweights outliers. However, the current construction assumes a single elliptical component per cluster, and the number of prototypes K is fixed a priori rather than inferred via Bayesian nonparametric priors. Future efforts will extend this conjugate construction to broader scale mixtures, other latent-variable generative models, and explore automatic model selection for dynamic prototype discovery in longitudinal image analysis.

Acknowledgments and Disclosure of Funding

This work was supported by the National Natural Science Foundation of China (Serial No. 62406182), and Fundamental Research Program of Shanxi Province(Serial No. 202403021212176).

References

[1] N. Abiri and M. Ohlsson. Variational auto-encoders with student's t-prior. In *27th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, 2019.

[2] D. F. Andrews and C. L. Mallows. Scale mixtures of normal distributions. *Journal of the Royal Statistical Society: Series B*, 36(1):99–102, 1974.

[3] D. Ba, B. Babadi, P. L. Purdon, and E. N. Brown. Convergence and stability of iteratively re-weighted least squares algorithms. *IEEE Transactions on Signal Processing*, 62(1):183–195, 2013.

[4] C. Blaiotta, P. Freund, M. J. Cardoso, and J. Ashburner. Generative diffeomorphic modelling of large MRI data sets for probabilistic template construction. *NeuroImage*, 166:117–134, 2018.

[5] B. Boenninghoff, S. Zeiler, R. Nickel, and D. Kolossa. Variational autoencoder with embedded Student-t mixture model for authorship attribution. In *Proceedings of the 28th International Conference on Computational Linguistics (COLING)*, pages 519–529, Barcelona, Spain (Online), 2020. International Committee on Computational Linguistics.

[6] M. Caron, P. Bojanowski, A. Joulin, and M. Douze. Deep clustering for unsupervised learning of visual features. In *European Conference on Computer Vision (ECCV)*, 2018.

[7] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[8] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin. Emerging properties in self-supervised vision transformers. In *IEEE International Conference on Computer Vision (ICCV)*, 2021.

[9] J. da Silva Gonçalves, L. Manduchi, M. Vandenhirtz, and J. E. Vogt. Treediffusion: Hierarchical generative clustering for conditional diffusion. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 447–462. Springer, 2025.

[10] A. Dalca, M. Rakic, J. Guttag, and M. Sabuncu. Learning conditional deformable templates with convolutional networks. *Advances in neural information processing systems*, 32, 2019.

[11] N. Dilokthanakul, P. A. M. Mediano, M. Garnelo, M. C. H. Lee, H. Salimbeni, K. Arulkumaran, and M. Shanahan. Deep unsupervised clustering with Gaussian mixture variational autoencoders. *arXiv preprint arXiv:1611.02648*, 2016.

[12] X. Guo, L. Gao, X. Liu, and J. Yin. Improved deep embedded clustering with local structure preservation. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 2017.

[13] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[14] Z. Jiang, Y. Zheng, H. Tan, B. Tang, and H. Zhou. Variational deep embedding: An unsupervised and generative approach to clustering. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 2017.

[15] J. Kim, J. Kwon, M. Cho, H. Lee, and J.-H. Won. t^3-variational autoencoder: Learning heavy-tailed data with student's t and power divergence. In *International Conference on Learning Representations*, volume 2024, pages 19580–19606, 2024.

[16] D. P. Kingma and M. Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[17] T. Kobayashi. Variational deep embedding with regularized student-t mixture model. In *International Conference on Artificial Neural Networks*, pages 443–455. Springer, 2019.

[18] P. J. LaMontagne, T. L. Benzinger, J. C. Morris, S. Keefe, R. Hornbeck, C. Xiong, E. Grant, J. Hassenstab, K. Moulder, A. G. Vlassenko, et al. Oasis-3: longitudinal neuroimaging, clinical, and cognitive dataset for normal aging and alzheimer disease. *medrxiv*, pages 2019–12, 2019.

[19] K. L. Lange, R. J. A. Little, and J. M. G. Taylor. Robust statistical modeling using the t distribution. *Journal of the American Statistical Association*, 84(408):881–896, 1989.

[20] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2332, 1998.

[21] Y. Li, P. Hu, Z. Liu, D. Peng, J. T. Zhou, and X. Peng. Contrastive clustering. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2021.

[22] G. J. McLachlan, S. X. Lee, and S. I. Rathnayake. Finite mixture models. *Annual review of statistics and its application*, 6(1):355–378, 2019.

[23] D. Peel and G. J. McLachlan. Robust mixture modelling using the t distribution. *Statistics and Computing*, 10(4):339–348, 2000.

[24] M. Rakic, A. Hoopes, M. S. Abulnaga, M. R. Sabuncu, J. V. Guttag, A. V. Dalca, A. D. N. Initiative, et al. Atlasmorph: Learning conditional deformable templates for brain mri. *Medical Image Analysis*, 110:103893, 2026.

[25] A. H. Song, R. J. Chen, T. Ding, D. F. Williamson, G. Jaume, and F. Mahmood. Morphological prototyping for unsupervised slide representation learning in computational pathology. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11566–11578, 2024.

[26] M. Svensén and C. M. Bishop. Robust bayesian mixture modelling. *Neurocomputing*, 64:235–252, 2005.

[27] W. Van Gansbeke, S. Vandenhende, S. Georgoulis, M. Proesmans, and L. Van Gool. Scan: Learning to classify images without labels. In *European conference on computer vision*, pages 268–285. Springer, 2020.

[28] J. Wang and M. Zhang. Bayesian atlas building with hierarchical priors for subject-specific regularization. In *Medical Image Computing and Computer Assisted Intervention (MICCAI)*. Springer, 2021.

[29] J. Wang and M. Zhang. Geo-sic: learning deformable geometric shapes in deep image classifiers. *Advances in Neural Information Processing Systems*, 35:27994–28007, 2022.

[30] Y. Wang, Y. Schiff, A. Gokaslan, W. Pan, F. Wang, C. De Sa, and V. Kuleshov. Infodiffusion: Representation learning using information maximizing diffusion models. In *International conference on machine learning*, pages 36336–36354. PMLR, 2023.

[31] H. Xia, H. Huang, and Z. Ding. Gcc: Generative calibration clustering. *arXiv preprint arXiv:2404.09115*, 2024.

[32] H. Xiao, K. Rasul, and R. Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

[33] J. Xie, R. Girshick, and A. Farhadi. Unsupervised deep embedding for clustering analysis. In *International Conference on Machine Learning (ICML)*, 2016.

[34] J. Xing, T. Zeevi, N. Wu, and J. Wang. On-manifold variational learning with heat-kernel priors. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 110–120. Springer, 2026.

[35] J. Yan, J. Liu, and Z.-y. Zhang. Clusterddpm: an em clustering framework with denoising diffusion probabilistic models. *Information Sciences*, page 122518, 2025.

Appendix

This appendix supplies the complete analytical derivations underlying the closed-form ELBO presented in the main paper (Sections A–E), and additional experimental results (Section F). The central claim of the paper, that the Student-t variant of VaDE admits an exact analytical objective with no approximation in any KL term, rests on three results: the scale-mixture identity that recasts the Student-t prior as a Gamma–Gaussian hierarchy (Section B), the Gamma–Gaussian conjugacy that determines the optimal scale posterior in closed form (Section C), and the analytical integration of the auxiliary variable that collapses the augmented ELBO into three tractable KL divergences (Section D). We prove each in turn and extend the result to the full-covariance setting (Section E). Section A collects the notation and the key equations from the main paper that serve as reference points for the derivations.

A Notation and Key Equations

For self-containedness, we restate the conventions and the principal equations from the main paper.

Notation. Observations $\boldsymbol{x} \in \mathbb{R}^D$ are embedded into latent codes $\boldsymbol{z} \in \mathbb{R}^d$ with cluster labels $c \in \{1, \dots, K\}$. We write $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1})$ for a Gaussian with mean $\boldsymbol{\mu}$ and precision $\boldsymbol{\Lambda}$, $\mathcal{T}_\nu(\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1})$ for a Student's t with ν degrees of freedom, and $\mathrm{Gamma}(\alpha, \beta)$ for a Gamma in shape–rate form (mean α/β). Cluster precisions $\boldsymbol{\Lambda}_c$ are diagonal in the main text with entries $\{\Lambda_{c,j}\}_{j=1}^d$; the full-covariance case is treated in Section E. The digamma function is ψ.

Generative model.

$$c \sim \mathrm{Cat}(\boldsymbol{\pi}), \quad u \mid c \sim \mathrm{Gamma}(\nu_c/2, \nu_c/2), \quad \boldsymbol{z} \mid u, c \sim \mathcal{N}(\boldsymbol{\mu}_c, (u\boldsymbol{\Lambda}_c)^{-1}), \quad \boldsymbol{x} \mid \boldsymbol{z} \sim p_\theta(\boldsymbol{x} \mid \boldsymbol{z}). \tag{A.1}$$

Variational family.

$$q(\boldsymbol{z}, u, c \mid \boldsymbol{x}) = q_\phi(\boldsymbol{z} \mid \boldsymbol{x})\, q(u \mid \boldsymbol{x}, c)\, q(c \mid \boldsymbol{x}), \tag{A.2}$$

with $q_\phi(\boldsymbol{z} \mid \boldsymbol{x}) = \mathcal{N}(\boldsymbol{\mu}_\phi(\boldsymbol{x}), \mathrm{diag}(\boldsymbol{\sigma}^2_\phi(\boldsymbol{x})))$ and $q(c \mid \boldsymbol{x}) = \gamma_c$.

Encoder-averaged Mahalanobis distance.

$$\bar{\Delta}^2_c \coloneqq \mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}\big[(\boldsymbol{z} - \boldsymbol{\mu}_c)^\top \boldsymbol{\Lambda}_c (\boldsymbol{z} - \boldsymbol{\mu}_c)\big] = \sum_{j=1}^{d} \Lambda_{c,j}\big[\sigma^2_{\phi,j} + (\mu_{\phi,j} - \mu_{c,j})^2\big]. \tag{A.3}$$

Optimal scale posterior.

$$q^\star(u \mid \boldsymbol{x}, c) = \mathrm{Gamma}(\tilde{\alpha}_c, \tilde{\beta}_c), \quad \tilde{\alpha}_c = \tfrac{\nu_c + d}{2}, \quad \tilde{\beta}_c = \tfrac{\nu_c + \bar{\Delta}^2_c}{2}. \tag{A.4}$$

Sufficient statistics.

$$\mathbb{E}[u] = \tfrac{\nu_c + d}{\nu_c + \bar{\Delta}^2_c}, \qquad \mathbb{E}[\log u] = \psi(\tilde{\alpha}_c) - \log \tilde{\beta}_c. \tag{A.5}$$

Collapsed ELBO and KL terms.

$$\mathcal{L}(\boldsymbol{x}) = \mathbb{E}_{q_\phi}[\log p_\theta(\boldsymbol{x} \mid \boldsymbol{z})] - \sum_{c=1}^{K} \gamma_c \big[\widetilde{\mathrm{KL}}^{(c)}_{\text{Gauss}} + \mathrm{KL}^{(c)}_{\text{Gamma}} + \mathrm{KL}^{(c)}_{\text{Cat}}\big], \tag{A.6}$$

$$\widetilde{\mathrm{KL}}^{(c)}_{\text{Gauss}} = -\tfrac{1}{2}\Big[d + d\,\mathbb{E}[\log u] + \log|\boldsymbol{\Lambda}_c| + \textstyle\sum_j \log \sigma^2_{\phi,j} - \mathbb{E}[u]\,\bar{\Delta}^2_c\Big], \tag{A.7}$$

$$\mathrm{KL}^{(c)}_{\text{Gamma}} = (\tilde{\alpha}_c - \alpha_0)\,\psi(\tilde{\alpha}_c) + \log\Gamma(\alpha_0) - \log\Gamma(\tilde{\alpha}_c) + \alpha_0 \log \tfrac{\tilde{\beta}_c}{\beta_0} + \tilde{\alpha}_c\, \tfrac{\beta_0 - \tilde{\beta}_c}{\tilde{\beta}_c}, \tag{A.8}$$

with prior $\alpha_0 = \beta_0 = \nu_c/2$, and $\mathrm{KL}^{(c)}_{\text{Cat}} = \log \gamma_c - \log \pi_c$.

B Proof of the Scale Mixture Representation

The starting point of our derivation is the observation that a Student-t density is itself an integral of Gaussians against a Gamma. This identity is what allows us to introduce the auxiliary variable u in (A.1) without changing the marginal distribution of $\boldsymbol{z}$, and is therefore the structural foundation on which the subsequent collapsed inference rests.

Proposition B.1 (Scale Mixture of Gaussians)**.** *If* $u \sim \mathrm{Gamma}(\nu/2, \nu/2)$ *and* $\boldsymbol{z} \mid u \sim \mathcal{N}(\boldsymbol{\mu}, (u\boldsymbol{\Lambda})^{-1})$*, then* $\boldsymbol{z} \sim \mathcal{T}_\nu(\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1})$.

Proof. The proof proceeds by direct integration of the joint density $p(\boldsymbol{z}, u) = p(\boldsymbol{z} \mid u)p(u)$ over u. The key step is recognising that the integrand is, up to constants, a Gamma kernel.

Step 1: joint density. The conditional Gaussian is $p(\boldsymbol{z} \mid u) = \frac{|u\boldsymbol{\Lambda}|^{1/2}}{(2\pi)^{d/2}} \exp\{-\frac{u}{2}\Delta^2\}$ with $\Delta^2 = (\boldsymbol{z}-\boldsymbol{\mu})^\top \boldsymbol{\Lambda}(\boldsymbol{z}-\boldsymbol{\mu})$ and $|u\boldsymbol{\Lambda}| = u^d|\boldsymbol{\Lambda}|$. The Gamma prior is $p(u) = \frac{(\nu/2)^{\nu/2}}{\Gamma(\nu/2)} u^{\nu/2-1} e^{-\nu u/2}$. Multiplying these and collecting powers of u produces an integrand proportional to $u^{(\nu+d)/2-1} \exp\{-\frac{u}{2}(\nu + \Delta^2)\}$, which is the kernel of a Gamma distribution.

Step 2: integration.

$$p(\boldsymbol{z}) = \frac{|\boldsymbol{\Lambda}|^{1/2}(\nu/2)^{\nu/2}}{(2\pi)^{d/2}\Gamma(\nu/2)} \int_0^\infty u^{(\nu+d)/2-1} \exp\big\{-\tfrac{u}{2}(\nu + \Delta^2)\big\}\, \mathrm{d}u. \tag{B.1}$$

The remaining integral is in standard form $\int_0^\infty t^{a-1}e^{-bt}\mathrm{d}t = \Gamma(a)/b^a$, with $a = (\nu + d)/2$ and $b = (\nu + \Delta^2)/2$, giving

$$\int_0^\infty u^{(\nu+d)/2-1} \exp\big\{-\tfrac{u}{2}(\nu + \Delta^2)\big\}\, \mathrm{d}u = \frac{\Gamma\big(\frac{\nu+d}{2}\big)}{\big(\frac{\nu+\Delta^2}{2}\big)^{(\nu+d)/2}}. \tag{B.2}$$

Step 3: simplification. Using $\big((\nu + \Delta^2)/2\big)^{(\nu+d)/2} = (\nu/2)^{(\nu+d)/2}(1 + \Delta^2/\nu)^{(\nu+d)/2}$, $(\nu/2)^{\nu/2}/(\nu/2)^{(\nu+d)/2} = (\nu/2)^{-d/2}$, and $(\nu/2)^{-d/2}/(2\pi)^{d/2} = (\nu\pi)^{-d/2}$, the constants combine into the standard Student-t normalising factor:

$$p(\boldsymbol{z}) = \frac{\Gamma\big(\frac{\nu+d}{2}\big)|\boldsymbol{\Lambda}|^{1/2}}{\Gamma\big(\frac{\nu}{2}\big)(\nu\pi)^{d/2}} \Big[1 + \tfrac{\Delta^2}{\nu}\Big]^{-(\nu+d)/2}, \tag{B.3}$$

which is the density of $\mathcal{T}_\nu(\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1})$. □

The identity has two consequences that the main paper relies on. First, the augmented model (A.1) is statistically indistinguishable from a t-mixture VaDE: every marginal distribution is unchanged. Second, conditional on u, the prior on $\boldsymbol{z}$ is Gaussian, which is precisely what makes the inference problem tractable in the next section.

C Optimal Scale Posterior

The structured factorisation (A.2) leaves the form of $q(u \mid \boldsymbol{x}, c)$ unspecified. We now show that this freedom is in fact illusory: the coordinate-ascent optimum is uniquely determined by Gamma–Gaussian conjugacy, and admits the closed-form expression quoted in (A.4). This result is what allows us to treat u as analytically integrable rather than approximated.

Proposition C.1 (Optimal scale posterior)**.** *Under the augmented model* (A.1) *and the variational family* (A.2)*, the optimal* $q(u \mid \boldsymbol{x}, c)$ *is the Gamma distribution given by* (A.4)*.*

Proof. The coordinate-ascent variational inference (CAVI) update for $q(u \mid \boldsymbol{z}, c)$ holding $q_\phi(\boldsymbol{z} \mid \boldsymbol{x})$ and $q(c \mid \boldsymbol{x})$ fixed is given by Bayes' rule restricted to the conditional model:

$$q(u \mid \boldsymbol{z}, c) \propto p(\boldsymbol{z} \mid u, c)\, p(u \mid c). \tag{C.1}$$

Substituting the conditional Gaussian and Gamma densities from (A.1),

$$q(u \mid \boldsymbol{z}, c) \propto u^{d/2} e^{-\frac{u}{2}\Delta_c^2(\boldsymbol{z})} \cdot u^{\nu_c/2-1} e^{-\frac{\nu_c}{2}u} = u^{(\nu_c+d)/2-1} e^{-\frac{u}{2}(\nu_c+\Delta_c^2(\boldsymbol{z}))}, \tag{C.2}$$

which is recognised as the kernel of $\mathrm{Gamma}\big((\nu_c+d)/2,\ (\nu_c+\Delta_c^2(\boldsymbol{z}))/2\big)$. Taking the expectation under the encoder $q_\phi(\boldsymbol{z} \mid \boldsymbol{x})$ replaces $\Delta_c^2(\boldsymbol{z})$ by its mean $\bar{\Delta}_c^2$ from (A.3), yielding $q^\star(u \mid \boldsymbol{x}, c) = \mathrm{Gamma}(\tilde{\alpha}_c, \tilde{\beta}_c)$ with the parameters in (A.4). □

The form of $\tilde{\alpha}_c$ and $\tilde{\beta}_c$ already foreshadows the robustness mechanism analysed in the main paper: an inlier with $\bar{\Delta}_c^2 \approx d$ produces a Gamma centred near the prior mean and hence $\mathbb{E}[u] \approx 1$, whereas an outlier with $\bar{\Delta}_c^2 \gg \nu_c$ produces a Gamma whose mass concentrates at small u, and hence $\mathbb{E}[u] \to 0$. Both expectations are required for the collapsed ELBO derived next, and both are available in closed form via (A.5).

D Collapsed ELBO Derivation

We now derive the collapsed ELBO (A.6). The derivation proceeds in three stages: (i) write the augmented ELBO under the variational family (A.2); (ii) regroup terms into three KL divergences corresponding to $\boldsymbol{z}$, u, and c; (iii) substitute the optimal $q^\star(u \mid \boldsymbol{x}, c)$ and integrate u analytically. Each KL term then reduces to a closed-form expression in the sufficient statistics (A.5).

Stage 1: augmented ELBO. By the definition of the ELBO under (A.1) and (A.2),

$$\begin{aligned}\mathcal{L} = \mathbb{E}_q\big[&\log p_\theta(\boldsymbol{x} \mid \boldsymbol{z}) + \log p(\boldsymbol{z} \mid u, c) + \log p(u \mid c) + \log p(c) \\ &- \log q_\phi(\boldsymbol{z} \mid \boldsymbol{x}) - \log q(u \mid \boldsymbol{z}, c) - \log q(c \mid \boldsymbol{x})\big].\end{aligned} \tag{D.1}$$

Stage 2: regrouping into three KL divergences. Pairing each prior term with its variational counterpart and using $q(c \mid \boldsymbol{x}) = \gamma_c$,

$$\mathcal{L} = \mathbb{E}_{q_\phi}[\log p_\theta(\boldsymbol{x} \mid \boldsymbol{z})] - \sum_c \gamma_c \left[\underbrace{\mathbb{E}_{q_\phi, q_u}\big[\log \tfrac{q_\phi(\boldsymbol{z}|\boldsymbol{x})}{p(\boldsymbol{z}|u,c)}\big]}_{\widetilde{\mathrm{KL}}^{(c)}_{\text{Gauss}}} + \underbrace{\mathbb{E}_{q_\phi}\big[\mathrm{KL}[q(u \mid \boldsymbol{z}, c) \| p(u \mid c)]\big]}_{\mathrm{KL}^{(c)}_{\text{Gamma}}} + \underbrace{\log \tfrac{\gamma_c}{\pi_c}}_{\mathrm{KL}^{(c)}_{\text{Cat}}} \right]. \tag{D.2}$$

Each of the three braced terms is now treated separately.

Stage 3a: collapsing the Gaussian KL. Substituting the Gaussian log-density $\log p(\boldsymbol{z} \mid u, c) = -\frac{d}{2}\log 2\pi + \frac{d}{2}\log u + \frac{1}{2}\log|\boldsymbol{\Lambda}_c| - \frac{u}{2}\Delta_c^2(\boldsymbol{z})$ and the encoder log-density into the first term, and using the four identities

$$\mathbb{E}_{q_u}[\log u] = \mathbb{E}[\log u], \quad \mathbb{E}_{q_u}[u] = \mathbb{E}[u], \quad \mathbb{E}_{q_\phi}[\Delta_c^2(\boldsymbol{z})] = \bar{\Delta}_c^2, \quad \mathbb{E}_{q_\phi}\Big[\sum_j \tfrac{(z_j - \mu_{\phi,j})^2}{\sigma_{\phi,j}^2}\Big] = d, \tag{D.3}$$

the entire u-dependence collapses into the two scalars $\mathbb{E}[u]$ and $\mathbb{E}[\log u]$:

$$\begin{aligned}\widetilde{\mathrm{KL}}^{(c)}_{\text{Gauss}} &= \big(-\tfrac{1}{2}\textstyle\sum_j \log \sigma_{\phi,j}^2 - \tfrac{d}{2}\big) - \big(\tfrac{d}{2}\mathbb{E}[\log u] + \tfrac{1}{2}\log|\boldsymbol{\Lambda}_c| - \tfrac{\mathbb{E}[u]}{2}\bar{\Delta}_c^2\big) \\ &= -\tfrac{1}{2}\big[d + d\,\mathbb{E}[\log u] + \log|\boldsymbol{\Lambda}_c| + \textstyle\sum_j \log \sigma_{\phi,j}^2 - \mathbb{E}[u]\,\bar{\Delta}_c^2\big],\end{aligned} \tag{D.4}$$

which is (A.7).

Stage 3b: closing the Gamma KL. The second term is the KL divergence between two Gamma distributions in shape–rate form, $q = \mathrm{Gamma}(\tilde{\alpha}, \tilde{\beta})$ and $p = \mathrm{Gamma}(\alpha_0, \beta_0)$. Using $\mathbb{E}_q[\log u] = \psi(\tilde{\alpha}) - \log\tilde{\beta}$ and $\mathbb{E}_q[u] = \tilde{\alpha}/\tilde{\beta}$ in the standard Gamma–Gamma KL formula,

$$\mathrm{KL}[q\|p] = (\tilde{\alpha} - \alpha_0)\,\psi(\tilde{\alpha}) + \log\Gamma(\alpha_0) - \log\Gamma(\tilde{\alpha}) + \alpha_0 \log\tfrac{\tilde{\beta}}{\beta_0} + \tilde{\alpha}\,\tfrac{\beta_0 - \tilde{\beta}}{\tilde{\beta}}, \tag{D.5}$$

which becomes (A.8) on substituting $\tilde{\alpha} = \tilde{\alpha}_c$, $\tilde{\beta} = \tilde{\beta}_c$, and $\alpha_0 = \beta_0 = \nu_c/2$.

Stage 3c: categorical KL. The third term is identical to that of Gaussian VaDE, $\mathrm{KL}_{\mathrm{Cat}}^{(c)} = \log \gamma_c - \log \pi_c$, since the augmentation does not alter the cluster prior.

Recovery of the Gaussian limit. As $\nu_c \to \infty$, the Gamma prior on u concentrates at 1, giving $\mathbb{E}[u] \to 1$ and $\mathbb{E}[\log u] \to 0$, while $\mathrm{KL}_{\mathrm{Gamma}}^{(c)} \to 0$. Substituting into the three KL expressions, $\widetilde{\mathrm{KL}}_{\mathrm{Gauss}}^{(c)}$ reduces to the standard Gaussian–Gaussian KL of VaDE and the Gamma term vanishes identically. The collapsed ELBO (A.6) therefore recovers the Gaussian VaDE objective as a strict limiting case.

Summary. Each KL term in (A.6) has been derived in closed form, without recourse to Wilson–Hilferty transformations, moment matching, or variational mixture bounds. The auxiliary variable u enters the final loss only through its first two moments $\mathbb{E}[u]$ and $\mathbb{E}[\log u]$, both available in closed form via (A.5), so no sampling of u is required at any point during training.

E Full-Covariance Extension

The main paper assumes diagonal cluster precisions $\mathbf{\Lambda}_c$ to match the diagonal Gaussian encoder. We show here that the derivation extends to the full-covariance setting without modification of any KL term, so that the closed-form result is not an artefact of the diagonal assumption.

For full $\mathbf{\Lambda}_c \in \mathbb{R}^{d\times d}$, the encoder-averaged Mahalanobis distance generalises to

$$\bar{\Delta}_c^2 = \mathrm{tr}(\mathbf{\Lambda}_c \mathbf{\Sigma}_\phi) + (\boldsymbol{\mu}_\phi - \boldsymbol{\mu}_c)^\top \mathbf{\Lambda}_c (\boldsymbol{\mu}_\phi - \boldsymbol{\mu}_c), \tag{E.1}$$

where $\mathbf{\Sigma}_\phi = \mathrm{diag}(\boldsymbol{\sigma}_\phi^2)$ for the diagonal encoder used in the main text (the result also holds for a full-covariance encoder, replacing $\mathbf{\Sigma}_\phi$ with the encoder covariance), and $\log|\mathbf{\Lambda}_c|$ is the log-determinant of the full precision matrix.

The derivations of the optimal scale posterior (A.4), the sufficient statistics (A.5), and the KL terms (A.7)–(A.8) are unchanged in form: each depends on $\mathbf{\Lambda}_c$ only through the scalar $\bar{\Delta}_c^2$ and the determinant $\log|\mathbf{\Lambda}_c|$. The Gamma–Gaussian conjugacy underlying Proposition C.1 holds for any positive-definite $\mathbf{\Lambda}_c$, and the Stage-3a collapse goes through with the substitution $\sum_j \log \sigma_{\phi,j}^2 \to \log|\mathbf{\Sigma}_\phi|$ for a full-covariance encoder. The closed-form character of the ELBO is therefore preserved.

F Related Experimental Results

F.1 Visualization

We present visualization results in Figure F.1 for different types of noise injection on the MNIST, Fashion MNIST, and OASIS datasets. We provide the detailed experiments of three 2-d views of OASIS in Figure F.2.

F.2 Computational Costs

We distinguish training-update cost from end-to-end cluster assignment. Training measurements use an NVIDIA RTX 5880 Ada GPU, batch size 32, and 101 timed optimizer updates per reported configuration. The two-phase baseline is timed in its second training phase. These measurements do not represent the total duration of 101 epochs.

For the reverse-inference benchmark, existing checkpoints are evaluated at Cauchy noise $\gamma = 0.14$ over three seeds. The encoded latent is forward-noised to the final diffusion time, followed by 100 deterministic DDIM reverse updates before mixture assignment. The original raw-latent assignment path is not used in this benchmark. VaDE and our method use their usual encode-and-assign paths. Latency includes encoding, reverse updates where applicable, and mixture assignment. We use batch size 256, discard five warm-up calls, and average 20 CUDA-event timing repetitions per seed. Training still samples one diffusion time per update; the network is not retrained for this timing experiment.

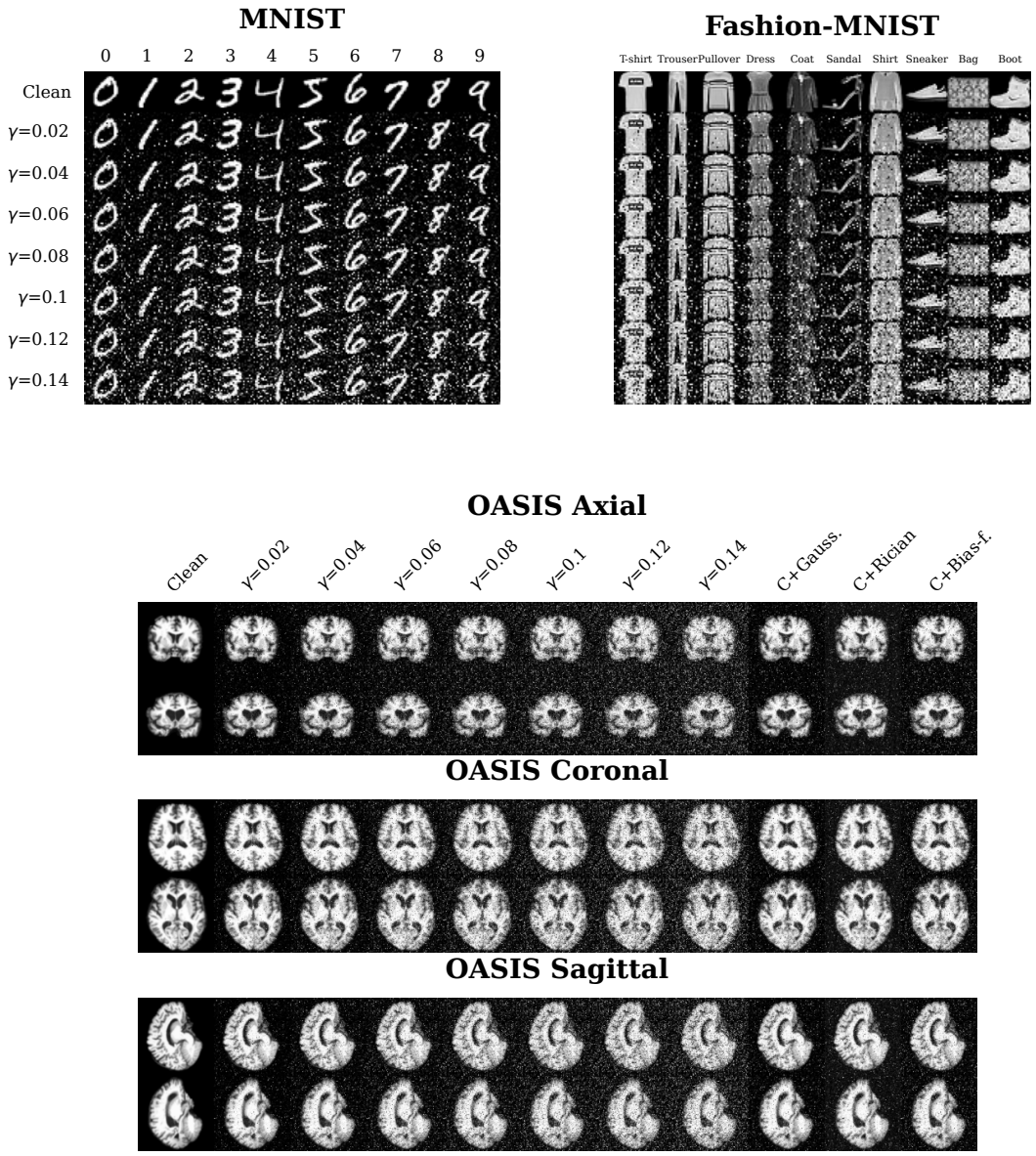
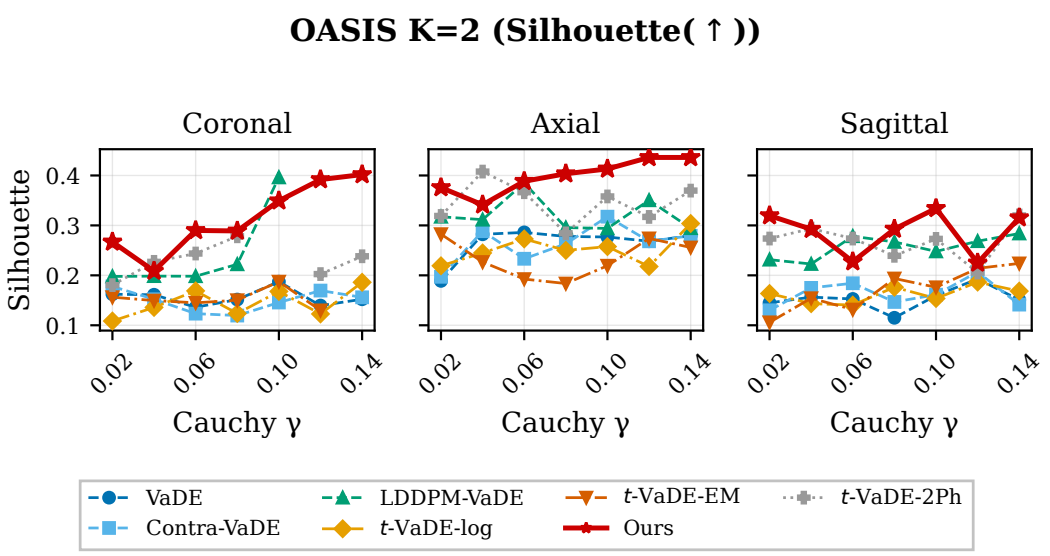
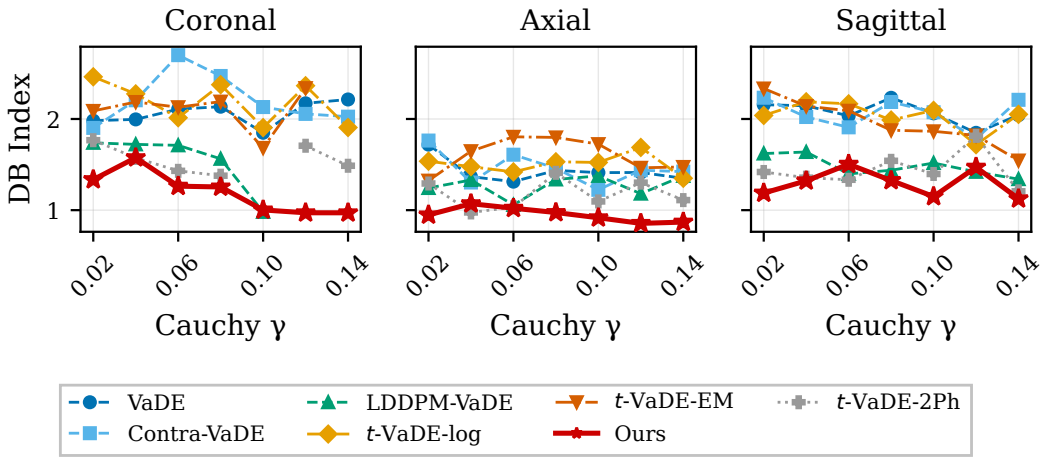


Figure F.1: Noisy input visualization across all benchmark datasets. **Top two panels:** MNIST and Fashion-MNIST with one sample per class (columns) under clean input and seven Cauchy noise levels $\gamma \in \{0.02, 0.04, \ldots, 0.14\}$ (rows). **Bottom three panels:** OASIS brain MRI in coronal, axial, and sagittal views (2 representative slices per panel). Each column corresponds to a noise condition: clean, seven Cauchy levels, and three composite types: Cauchy+Gaussian, Cauchy+Rician, and Cauchy+Bias-field at medium strength. Cauchy noise produces heavy-tailed outlier pixels that progressively corrupt local structure while global anatomy remains recognizable.



(a) Silhouette score across three views.



(b) Davies–Bouldin index across three views.

Figure F.2: OASIS (K=2) clustering quality under Cauchy noise for coronal, axial, and sagittal views, respectively. Our method consistently achieves the highest Silhouette score and lowest Davies–Bouldin index across all three anatomical views and noise levels (consistent with the results in Table 1 in the main paper), demonstrating robust clustering quality regardless of the imaging plane.

Table F.1: Controlled training cost in ms/update.

Method	MNIST	Fashion-MNIST
VaDE	27.74	27.05
Contra-VaDE	28.79	29.14
LDDPM-VaDE	31.99	30.20
t-VaDE-log	28.48	30.89
t-VaDE-EM	29.14	30.00
Ours	33.21	32.00
t-VaDE-2Ph	24.53	25.66

Table F.2: End-to-end assignment latency in ms/sample. The $33\times/36\times$ comparison is against the specified 100-step reverse-inference path.

Method	Reverse calls	MNIST	Fashion	Params (MB)
VaDE	0	0.0036	0.0036	6.95
LDDPM-VaDE-100	100	0.1422	0.1422	6.98
Ours	0	0.0043	0.0040	6.95

F.3 Diagonal versus Full Precision

We compare diagonal and full cluster precision matrices at Cauchy noise $\gamma = 0.14$ over three paired seeds, while retaining the same encoder and other settings. Full precision preserves the analytic bound and increases computation from $O(BKd)$ to $O(BKd^2)$ for the mixture terms.

Table F.3: Diagonal/full precision ablation at Cauchy $\gamma = 0.14$. ACC is mean$\pm$sample SD across three matched seeds. Time is measured in ms/update.

Dataset	Diag. ACC	Full ACC	Diag. time	Full time
MNIST	$.796 \pm .044$	$.812 \pm .075$	33.21	34.22
Fashion-MNIST	$.543 \pm .007$	$.568 \pm .016$	32.00	33.66

The full model therefore provides an analytic option when cross-dimensional dependencies are useful, while diagonal precision remains the lower-cost default.

F.4 Multi-Seed Results and Significance

We performance experiments over five random seeds at three Cauchy-noise scales. The mean paired difference on MNIST is 0.079 (95% CI [0.052, 0.106]) against t-VaDE-log; on Fashion-MNIST it is 0.116 (95% CI [0.032, 0.200]) against t-VaDE-EM. Paired t-tests with Holm correction across these two dataset-level comparisons yield $p = 0.0025$ and $p = 0.018$.

Table F.4: MNIST ACC over five seeds. Clean inputs are excluded.

Method	$\gamma = .02$	$\gamma = .08$	$\gamma = .14$
VaDE	$.663 \pm .051$	$.610 \pm .055$	$.622 \pm .042$
Contra-VaDE	$.694 \pm .041$	$.645 \pm .049$	$.606 \pm .035$
LDDPM-VaDE	$.328 \pm .096$	$.272 \pm .102$	$.197 \pm .070$
t-VaDE-log	$.749 \pm .022$	$.713 \pm .030$	$.687 \pm .037$
t-VaDE-EM	$.636 \pm .030$	$.627 \pm .051$	$.659 \pm .040$
t-VaDE-2Ph	$.652 \pm .061$	$.697 \pm .045$	$.625 \pm .068$
Ours	$.817 \pm .045$	$.778 \pm .010$	$.792 \pm .037$

Table F.5: Fashion-MNIST ACC over five seeds. Clean inputs are excluded.

Method	$\gamma = .02$	$\gamma = .08$	$\gamma = .14$
VaDE	$.455 \pm .136$	$.424 \pm .123$	$.311 \pm .131$
Contra-VaDE	$.479 \pm .028$	$.380 \pm .153$	$.356 \pm .070$
LDDPM-VaDE	$.191 \pm .063$	$.237 \pm .077$	$.185 \pm .027$
t-VaDE-log	$.502 \pm .025$	$.407 \pm .119$	$.363 \pm .117$
t-VaDE-EM	$.505 \pm .026$	$.483 \pm .074$	$.317 \pm .168$
t-VaDE-2Ph	$.516 \pm .023$	$.415 \pm .144$	$.261 \pm .205$
Ours	$.552 \pm .026$	$.555 \pm .019$	$.548 \pm .010$

F.5 Degrees-of-Freedom Initialization

We test initialization of $\nu_0 \in \{2, 4, 5, 8, 16, 32\}$ and evaluate three runs per setting at Cauchy $\gamma = 0.14$. Entries are mean$\pm$sample SD.

Table F.6: Sensitivity to the initial degrees of freedom.

ν_0	2	4	5	8	16	32
MNIST	$.830 \pm .044$	$.835 \pm .073$	$.796 \pm .044$	$.852 \pm .049$	$.803 \pm .015$	$.768 \pm .039$
Fashion	$.548 \pm .023$	$.544 \pm .009$	$.543 \pm .007$	$.536 \pm .027$	$.562 \pm .017$	$.551 \pm .050$

The Fashion-MNIST means range from 0.536 to 0.562. MNIST varies more, from 0.768 to 0.852, with the lowest mean at $\nu_0 = 32$.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: The abstract and introduction state four claims: (i) a fully closed-form ELBO for a Student's t-mixture deep clustering model, (ii) inherent noise immunity through implicit IRLS reweighting, (iii) state-of-the-art clustering accuracy and stability on standard benchmarks, and (iv) clinical utility on a real neurodegenerative disease cohort.

2. **Limitations**

   Answer: [Yes]

   Justification: Section 5 discusses the elliptical component geometry and the fixed number of components K. Appendix F.5 reports sensitivity to the initial degrees of freedom.

3. **Theory assumptions and proofs**

   Answer: [Yes]

   Justification: All theoretical claims are stated with explicit assumptions in Section 3. The closed-form ELBO (10) is derived from the structured variational posterior (6) together with the optimal $q^\star(u \mid \boldsymbol{x}, c)$ given by (8), with all expectation identities and the extension to full-covariance $\boldsymbol{\Lambda}_c$ provided in Appendix.

4. **Experimental result reproducibility**

   Answer: [Yes]

   Justification: Section 4.1 describes the encoder–decoder architecture, optimizer, learning rate, batch size, number of training epochs, initialization scheme for $(\boldsymbol{\mu}_c, \boldsymbol{\Lambda}_c, \pi_c, \nu_c)$, and the noise corruption protocol. Appendix F provides supplementary experimental results and timing protocols; training and evaluation scripts are available in the public code repository.

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

   Answer: [Yes]

   Justification: Our source code is publicly available at `https://github.com/xuzhikangnba/torch-tvae`. The repository includes the Student's t VAE implementation, the original multi-component models, training and evaluation scripts, installation instructions, and reproduction reports. MNIST and Fashion-MNIST can be downloaded through the provided training pipeline. OASIS data must be obtained separately from the dataset provider under its access conditions.

6. **Experimental setting/details**

   Answer: [Yes]

   Justification: Section 4.1 specifies the shared backbone, AdamW optimizer, learning rate, weight decay, batch size, training duration, and softplus parameterization of ν_c. Appendix F.5 reports the initialization sensitivity study.

7. **Experiment statistical significance**

   Answer: [Yes]

   Justification: Section 4.2 reports an additional five-seed comparison on MNIST and Fashion-MNIST, including 95% confidence intervals and p-values. This analysis is separate from the main results aggregated over noise levels or anatomical views. Table 1(b) additionally reports cluster-collapse counts and rates.

8. **Experiments compute resources**

   Answer: [Yes]

   Justification: All experiments were conducted on a workstation with four NVIDIA RTX 5880 Ada GPUs, each with 48 GB memory. Each individual training run used one GPU. The main experiments follow the 100-epoch configuration described in Section 4.1.

9. **Code of ethics**

Answer: [Yes]

Justification: The research conforms with the NeurIPS Code of Ethics. We use only publicly available datasets (MNIST, Fashion-MNIST, OASIS) under their respective licenses, and the OASIS Brain MRI dataset is used in compliance with its data use agreement.

10. **Broader impacts**

    Answer: [Yes]

    Justification: Our method enables more reliable identification of disease subtypes from medical imaging in the presence of noise, with potential positive impact on clinical analysis of disease heterogeneity (e.g., discovering rare subtypes that conventional models would merge). On the negative side, robust unsupervised clustering applied to medical data carries the standard risks of misinterpretation if used without clinical oversight, and like any method trained on retrospective cohorts may inherit demographic biases present in the source data; clinical deployment would require careful validation by domain experts and bias audits.

11. **Safeguards**

    Answer: [N/A]

    Justification: The released assets consist of research code, documentation, and numerical experiment records for image prototype learning. No pretrained model weights, patient images, or subject-level clinical records are distributed. MRI data must be obtained separately through the dataset provider under its access conditions.

12. **Licenses for existing assets**

    Answer: [Yes]

    Justification: All datasets and baseline implementations used in this paper are properly cited. MNIST [20] is released under Creative Commons Attribution-Share Alike 3.0 license; Fashion-MNIST [32] is released under MIT license; OASIS Brain MRI [18] is used under its publicly available data use agreement. Baseline implementations of VaDE [14], Contra-VaDE [21], ClusterDDPM [35], t^3VAE [15], and the t-VaDE variants [17] are cited in Section 4.1, and we re-implement them following the published descriptions to ensure architectural fairness.

13. **New assets**

    Answer: [Yes]

    Justification: We release an MIT-licensed Python package implementing a reusable Gaussian–Gamma prior, a basic label-free Student's t VAE, and the original multi-component models used by our method. The repository provides usage examples, training and evaluation scripts, configuration documentation, numerical audit results, and fresh reproduction reports. The documentation distinguishes the basic single-component model from the paper's clustering models and states the scope and limitations of the available reproduction results. The source code is available at `https://github.com/xuzhikangnba/torch-tvae`.

14. **Crowdsourcing and research with human subjects**

    Answer: [N/A]

    Justification: This paper does not involve crowdsourcing or new data collection from human subjects. The OASIS Brain MRI cohort is a pre-existing publicly released dataset; we do not perform new recruitment or data collection.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Answer: [N/A]

    Justification: We use the OASIS Brain MRI dataset, which was collected under its own IRB-approved protocols by the original data providers, in accordance with their publicly available data use agreement. No new human subject research is conducted in this paper.

16. **Declaration of LLM usage**

    Answer: [N/A]

    Justification: LLMs are not part of the core methodology of this work. They were used only as writing aids (grammar, phrasing) and do not impact the scientific rigor, originality, or experimental claims.