Heavy-Tailed Class-Conditional Priors for Long-Tailed Generative Modeling

Anonymous authors
Paper under double-blind review

Abstract

Variational Autoencoders (VAEs) with global priors trained under an imbalanced empirical class distribution can lead to underrepresentation of tail classes in the latent space. While t^3VAE improves robustness via heavy-tailed Student's t-distribution priors, its single global prior still allocates mass proportionally to class frequency. We address this latent geometric bias by introducing C-t^3VAE, which assigns a per-class Student's t joint prior over latent and output variables. This design promotes uniform prior mass across class-conditioned components. To optimize our model we derive a closed-form objective from the γ-power divergence, and we introduce an equal-weight latent mixture for class-balanced generation. On SVHN-LT, CIFAR100-LT, and CelebA datasets, C-t^3VAE consistently attains lower FID scores than t^3VAE and Gaussian-based VAE baselines under severe class imbalance while remaining competitive in balanced or mildly imbalanced settings. In per-class F1 evaluations, our model outperforms the conditional Gaussian VAE across highly imbalanced settings. Moreover, we identify the mild imbalance threshold $\rho < 5$, for which Gaussian-based models remain competitive. However, for $\rho \geq 5$ our approach yields improved class-balanced generation and mode coverage.

1 Introduction

Class imbalance and long-tail distributions are common in real-world datasets, yet generative models often fail to represent rare classes accurately. Under skewed training data, models tend to overfit dominant modes and underrepresent minority ones in latent and output spaces, resulting in biased generations. This issue is particularly consequential in applications such as facial synthesis (Mehta et al., 2024) and medical imaging (Pinaya et al., 2022), where such biases can exacerbate social and diagnostic disparities (Naik & Nushi, 2023). Addressing class imbalance to ensure balanced representational capacity across categories remains a major challenge for all generative models.

For Generative Adversarial Networks (GANs) (Goodfellow et al., 2014), their inherently unstable training dynamics, exacerbated by data imbalance, often lead to biased and mode-collapsed generations. Several works have sought to mitigate these effects. The Wasserstein GAN (Arjovsky et al., 2017) replaces the Jensen–Shannon divergence with the Wasserstein distance to stabilize optimization and improve sample diversity. PacGAN (Lin et al., 2018) enhances robustness by packing multiple samples into the discriminator, reducing mode collapse and improving diversity under implicit imbalances. Similarly, (Asokan & Seelamantula, 2020) introduces negative data augmentation to prevent under-representation of minority classes, enabling more class-balanced supervised GAN training. More recently, RareGAN (Lin et al., 2022) addresses unlabeled long-tailed data through a weighted loss and adaptive labeling budgets, improving both balance and generative diversity. Despite these advances, GANs remain difficult to train reliably, especially under strong imbalance.

Diffusion models have emerged as a more stable alternative to GANs, achieving superior image quality and convergence behavior (Dhariwal & Nichol, 2021). Several diffusion-based methods explicitly tackle data imbalance. Class-Balancing Diffusion Models (Qin et al., 2023) modify the denoising process to be class-invariant, while (Zhang et al., 2024) employs weighted score matching with Bayesian calibration to

transfer knowledge from majority to minority classes, improving diversity and fidelity on long-tailed datasets. Heavy-Tailed Diffusion Models (Pandey et al., 2025) take a different approach by replacing the Gaussian noise assumption with a Student's t formulation, offering a more robust fit for imbalanced distributions. Nevertheless, effectively addressing class imbalance within diffusion frameworks remains an open and active research area.

Variational Autoencoders (VAEs) (Kingma & Welling, 2013) are valuable models when interpretable, structured latent spaces and efficient single-step sampling are needed. VAEs provide a principled probabilistic foundation, stable training dynamics, and compatibility with latent-variable pipelines such as latent diffusion (Rombach et al., 2022). However, in class imbalanced setting, standard VAEs using isotropic Gaussian priors struggle to capture heavy-tailed or rare structures, creating class-coverage bottlenecks (Tam & Dunson, 2025). Therefore, prior work explores non-Gaussian priors, especially Student's t-distributions (Takahashi et al., 2018; Abiri & Ohlsson, 2020; Eguchi, 2021; Kim et al., 2024), to improve robustness to outliers and class imbalance.

In this work, we view class imbalance not only as a sampling issue but as a geometric one: optimizing under the empirical distribution tends to allocate latent probability mass in proportion to class frequency. Consequently, majority classes occupy larger effective regions of latent space, while minority classes are confined to smaller regions with reduced representational capacity. This interpretation is particularly transparent in latent-variable models with an explicit global prior, such as VAEs, where representational allocation is directly governed by prior density. While imbalance also affects GANs and diffusion models, in those frameworks it is intertwined with adversarial training dynamics or iterative score-based denoising. In contrast, VAEs allow us to study how prior specification alone shapes latent geometry. We therefore focus on VAEs to isolate the role of prior-induced frequency bias in a controlled and analytically tractable setting.

Current VAE based approaches retain a single global prior (Kingma & Welling, 2013; Takahashi et al., 2018; Mathieu et al., 2019b; Abiri & Ohlsson, 2020). Hence through our geometric view the latent remains frequency-aligned where class regions scale with empirical probability and heavy tails alone do not eliminate this frequency-induced geometric bias. Indeed, optimizing a VAE under $p_{data}(x) = \sum_y p(y)p(x|y)$ tends to allocate latent regions whose effective volume correlates with class frequency $p(y)$. We address this issue by introducing the Conditional-t^3VAE (C-t^3VAE), which imposes a per-class Student's t-distribution prior over the joint latent-output space. The design promotes uniform prior mass across class-conditioned components under the sampling distribution, mitigating majority-class dominance while the heavy tails capture intra-class variation. For class-balanced generation, we derive an equal-weight latent mixture of Student's t-distributions with analytically derived component variances. Importantly, C-t^3VAE class-specific heavy-tailed priors induce a structured latent mixture whose geometry differs fundamentally from both global heavy-tailed priors and Gaussian conditional VAEs. We summarize our main contributions in the following points :

- We propose the C-t^3VAE model with a training objective based on the γ-power divergence.
- We develop an equal-weight latent mixture sampling scheme with analytically derived optimal variance scaling for each component.
- We outperform relevant baselines in FID on SVHN-LT (Netzer et al., 2011), CIFAR100-LT (Cao et al., 2019), and CelebA (Liu et al., 2015) under severe imbalance, and show via per-class evaluation that C-t^3VAE better avoids mode collapse, exceeding a Gaussian conditional VAE in per class Recall and F1 while remaining competitive on Precision.
- We empirically observe a transition regime around $\rho \approx 5$, beyond which Gaussian priors become suboptimal, providing guidance for model selection on skewed datasets.

2 Related works

Since the introduction of Variational Autoencoders (VAEs) (Kingma & Welling, 2013), numerous extensions have aimed to enhance latent representation expressiveness by replacing the standard Gaussian prior

with more flexible formulations. Notable examples include Gaussian mixtures (Dilokthanakul et al., 2016; Saseendran et al., 2021), hyperspherical priors (Davidson et al., 2018), normalizing flows (Jaini et al., 2020), Riemannian manifolds (Chadebec et al., 2023), and implicit distributions (Takahashi et al., 2019). Most of these retain the Evidence Lower Bound (ELBO) optimization framework, while others introduce alternative divergence measures for greater modeling flexibility (Makhzani et al., 2016). Hierarchical VAEs (Vahdat & Kautz, 2020) and vector-quantized VAEs (van den Oord et al., 2017) have also been proposed to improve disentanglement and mitigate posterior collapse.

Student's t-distributions have been widely studied for their heavy-tailed robustness to outliers (Tam & Dunson, 2025). Early works (Mathieu et al., 2019a; Abiri & Ohlsson, 2020) incorporated t-distributed priors into the VAE framework using KL-divergence-based ELBO objectives to promote robust latent encodings. However, because the KL divergence between t-distributions lacks a closed-form solution, such approaches depend on numerical approximations, increasing computational cost. (Takahashi et al., 2018) proposes a Student's t decoder to improve generative performance, but the heavy-tailed assumption is confined to the decoder, leaving the latent space Gaussian and limiting validation to tabular data. The t^3VAE (Kim et al., 2024) further advances this line of work by jointly modeling latent and output distributions under Student's t-assumptions for the encoder, decoder, and prior, replacing the KL divergence with a closed-form γ-divergence (Eguchi, 2021).

While several works improve prior expressiveness or robustness, they do not explicitly analyze how global prior formulations induce frequency-aligned allocation of latent probability mass under imbalanced data. In particular, existing approaches focus on better density modeling or optimization stability, rather than controlling how representational capacity is distributed across classes. As a result, even expressive or heavy-tailed priors may continue to reflect empirical class proportions in latent space.

A complementary line of research addresses imbalance through conditional modeling. Conditional VAEs (CVAEs) (Kingma et al., 2014; Sohn et al., 2015) condition latent encodings on class labels, enabling targeted generation of minority categories. However, their Gaussian priors poorly approximate heavy-tailed or rare data structures. Our C-t^3VAE bridges two threads: it inherits the heavy-tailed robustness of t^3VAE while adopting the class-conditional structure of CVAEs, and further introduces a theoretically grounded equal-weight mixture sampling scheme that explicitly counteracts majority-class dominance at generation time. Hence, our contribution is not merely heavy tails nor conditioning, but decoupling representational capacity from empirical frequency through prior mass allocation.

3 Background

This section introduces the theoretical background and baseline models relevant to our work. We assume access to a labeled, imbalanced dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$, where $x_i \in \mathbb{R}^n$ is a data sample of dimension n, $y_i \in \{1, \dots, K\}$ its class label with K being the number of classes and m being the latent space dimension.

3.1 VAEs and Conditional VAEs

VAEs (Kingma & Welling, 2013) are generative models trained via variational inference by maximizing the Evidence Lower Bound (ELBO) of the log-likelihood. The standard objective of this model is

$$\mathcal{L}_{\theta,\phi} := \mathbb{E}_{z\sim q_\phi(\cdot|x)}[\log p_\theta(x|z)] - \mathcal{D}_{KL}(q_\phi(z|x)\|p(z)), \tag{1}$$

where the first term is the reconstruction loss with $p_\theta(x|z)$ being the decoder model. The second term is the Kullback–Leibler (KL) divergence between the approximate posterior $q_\phi(z|x)$ and the prior $p(z)$. The β-VAE is a weighted variant of the VAE model which introduces a β scaling term for the KL divergence (Higgins et al., 2017):

$$\mathcal{L}_{\theta,\phi} := \mathbb{E}_{z\sim q_\phi(\cdot|x)}[\log p_\theta(x|z)] - \beta\mathcal{D}_{KL}(q_\phi(z|x)\|p(z)), \tag{2}$$

with $p(z) \sim \mathcal{N}_m(0, I)$, $q_\phi(\cdot|x) \sim \mathcal{N}_m(\mu_\phi(x), \Sigma_\phi(x))$, and $p_\theta(x|z) \sim \mathcal{N}_m(\mu_\theta(z), \sigma^2 I)$. $\mu_\phi(\cdot)$ and $\Sigma_\phi(\cdot)$ are the mean and covariance matrices of the approximate posterior. They are inferred by a neural network with parameters ϕ given the input x. Moreover, $\mu_\theta(\cdot)$ is the decoder neural network with parameter θ and σ is a

parameter controlling the decoder's output covariance. This variant of the VAE model allows to place more weight on disentangling the latent space or on the reconstruction of the data points. To generate samples from the VAE or the β-VAE model, we sample a latent vector $z \sim \mathcal{N}_m(0, I)$. Then, the generated data point would be $\hat{x} \sim \mathcal{N}_m(\mu_\theta(z), \sigma^2 I)$.

Nevertheless, since Eq. (1) and Eq. (2) optimize the ELBO over the data distribution $p_{data}(x)$, which can be decomposed as $p_{\text{data}}(x) = \sum_{y_i} p(y_i)\, p_{\text{data}}(x \mid y_i)$. In the context of imbalanced data, this optimization inherently biases the model toward head classes with larger $p(y_i)$. As a result, most generated samples come from overrepresented classes, while tail classes' samples are underrepresented and of lower quality, a phenomenon commonly referred to as *mode collapse*. Therefore, when labels are available, it is preferable to define class-conditional posterior and prior distributions: $q_\phi(z|x, y)$ and $p(z|y)$. This yields the Conditional-VAE (CVAE) model trained using the objective (Kingma et al., 2014):

$$\sum_y \mathbb{E}_{z\sim q_\phi(\cdot|x,y)}[\log p_\theta(x|z, y)] - \beta \mathcal{D}_{KL}(q_\phi(z|x, y)\|p(z|y)). \tag{3}$$

In practice, in Eq. (3) we optimize the class-conditional objective with uniform weighting across labels, removing explicit frequency-dependent scaling from the loss. Also, we define $p(z|y) \sim \mathcal{N}_m(\mu_y, I)$ with learnable class-wise means μ_y. To generate a data point $\hat{x}_y$ from class y, we sample $z_y \sim \mathcal{N}_m(\mu_y, I)$, then we get $\hat{x}_y \sim p_\theta(x|z_y, y)$. Nevertheless, despite conditioning, this formulation remains Gaussian. Unlike Student's t-distributions, Gaussian priors poorly approximate heavy-tailed data distributions (Tam & Dunson, 2025).

3.2 Multivariate Student's t-Distribution

A d-dimensional Student's t-distribution with mean $\mu \in \mathbb{R}^d$, covariance $\Sigma \in \mathbb{R}^{d\times d}$, and degrees of freedom $\nu > 2$ is a heavy-tail, super-Gaussian distribution defined as

$$t_d(x) = C_{\nu,d}|\Sigma|^{-\frac{1}{2}} \left(1 + \frac{(x-\mu)^\top \Sigma^{-1}(x-\mu)}{\nu}\right)^{-\frac{\nu+d}{2}}, \qquad C_{\nu,d} = \frac{\Gamma\left(\frac{\nu+d}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)(\nu\pi)^{\frac{d}{2}}}. \tag{4}$$

The power form of this distribution prevents a closed-form KL divergence between two Student's t-distributions. Instead, the γ-power divergence $\mathcal{D}_\gamma(q\|p)$ is used (Eguchi, 2021; Kim et al., 2024). This divergence is defined for $q \sim t_d(\mu_0; \Sigma_0; \nu)$, $p \sim t_d(\mu_1; \Sigma_1; \nu)$ as

$$\mathcal{D}_\gamma(q\|p) := \frac{\mathcal{C}_\gamma(q, p) - \mathcal{H}_\gamma(p)}{\gamma} \tag{5}$$

with $\gamma = -\frac{2}{\nu+d}$, the γ-entropy $\mathcal{H}_\gamma(p)$ and γ-cross-entropy $\mathcal{C}_\gamma(q, p)$ being

$$\mathcal{H}_\gamma(p) := -\|p\|_{1+\gamma} = -\left(\int p(x)^{1+\gamma} dx\right)^{\frac{1}{1+\gamma}}, \qquad \mathcal{C}_\gamma(q, p) := -\int q(x)\left(\frac{p(x)}{\|p\|_{1+\gamma}}\right)^\gamma dx.$$

Then, substituting the definition of a Student's t-distribution from Eq. (4) into Eq. (5), the following closed-form formula for the γ-power divergence can be derived (Derivation of Proposition 3 in (Kim et al., 2024))

$$\begin{aligned}\mathcal{D}_\gamma(q\|p) = -\frac{C_{\nu,d}^{\frac{\gamma}{1+\gamma}}}{\gamma}\left(1 + \frac{d}{\nu-2}\right)^{-\frac{\gamma}{1+\gamma}} \Bigg[&- |\Sigma_0|^{-\frac{\gamma}{2(1+\gamma)}} \quad \left(1 + \frac{d}{\nu-2}\right) \\ &+ |\Sigma_1|^{-\frac{\gamma}{2(1+\gamma)}}\left(1 + \frac{\mathrm{Tr}\left(\Sigma_1^{-1}\Sigma_0\right)}{\nu-2} + \frac{(\mu_0-\mu_1)^\top \Sigma_1^{-1}(\mu_0-\mu_1)}{\nu}\right)\Bigg].\end{aligned} \tag{6}$$

3.3 t^3-Variational Autoencoder

3.3.1 Definition

The t^3VAE model (Kim et al., 2024) is a non-ELBO-based autoencoder which models the joint prior distribution $p_\theta(x, z)$ using multivariate Student's t-distribution

$$p_\theta(x,z) = \sigma^{-n} C_{\nu,m+n} \left[1 + \frac{1}{\nu}\left(\|z\|^2 + \frac{\|x-\mu_\theta(z)\|^2}{\sigma^2}\right)\right]^{-\frac{\nu+m+n}{2}}.$$

From this joint distribution, the marginal latent prior $p(z)$ and decoder distribution $p_\theta(x|z)$ can be defined as follows

$$p(z) = t_m(z|0, I, \nu), \quad p_\theta(x|z) = t_n\left(x\middle|\mu_\theta(z), \frac{1+\nu^{-1}\|z\|^2}{1+\nu^{-1}m}\sigma^2 I, \nu+m\right).$$

Furthermore, the posterior distribution is defined as :

$$q_\phi(z|x) = t_m\left(x\middle|\mu_\phi(x), \frac{\Sigma_\phi(x)}{1+\nu^{-1}n}, \nu+n\right).$$

Hence, the data distribution would be $q_\phi(x, z) = p_{\text{data}}(x)q_\phi(z|x)$. As a result, relying on the γ-divergence in Eq. (6) applied to the $p_\theta(x, z)$ and $q_\phi(x, z)$ distributions, the following loss function is derived to optimize the t^3VAE's parameters :

$$\mathcal{L}_\gamma = \mathbb{E}_x\left[\frac{\mathbb{E}_z\left[\|x-\mu_\theta(z)\|^2\right]}{\sigma^2} + \|\mu_\phi(x)\|^2 + \frac{\nu\,\mathrm{Tr}\left(\Sigma_\phi(x)\right)}{\nu+n-2} - \frac{\nu C_1}{C_2}|\Sigma_\phi(x)|^{-\frac{\gamma}{2(1+\gamma)}}\right], \tag{7}$$

with $\gamma = -\frac{2}{\nu+n+m}$ and C_1 and C_2 being constants theoretically derived in (Kim et al., 2024). We note that the first term in this loss function represents the standard reconstruction term in VAE models and the rest of the terms are regularization terms over the latent space. To sample from the latent space of the t^3VAE, (Kim et al., 2024) propose the $p^\star_\nu(z) = t_m(0, \tau^2 I, \nu+n)$ distribution with

$$\tau^2 = \left(1+\nu^{-1}n\right)^{-1}\left(\sigma^n C^{-1}_{\nu,n} \quad \frac{\nu+n-2}{\nu-2}\right)^{-\frac{2}{\nu+n-2}}. \tag{8}$$

Moreover, sampling from a multi-variate Student's t-distribution $T \sim t_d(\mu, \Sigma, \nu)$ both in the learning (Eq. (7)) and sampling (Eq. (8)) phases is performed through the standard reparameteration trick for Student's t-distributions $T := \mu + Z\sqrt{\nu V^{-1}}$ where $Z \sim \mathcal{N}(0, \Sigma)$ and $V \sim \mathcal{X}^2(\nu)$.

3.3.2 β-t^3VAE extension

From Eq. (7) we can also define a β-t^3VAE model by multiplying all the regularization terms by a β factor. Similarly to β-VAE models, this improves the versatility of the model and allows either a focus on generation or disentangling.

In summary, although the t^3VAE effectively models heavy-tailed distributions through Student's t-distributions and γ-power divergence, it does not explicitly address class imbalance in the latent space as it does not allocate equal prior mass across class-conditioned components. In the next section, we introduce a class-conditional variant of the t^3VAE, designed to enable class-balanced generation across all classes.

4 Conditional t^3-Variational Autoencoder

In this section, we propose the Conditional t^3-Variational Autoencoder (C-t^3VAE), present its formulation, training objective, and sampling strategy. C-t^3VAE models the latent space as a mixture of Student's t-distributions, one per class, inducing uniform prior mass across class-conditioned components under the sampling distribution. Intra-class variability is captured through the heavy-tailed nature of the Student's t prior.

4.1 Model definition

The C-t^3VAE we propose is based on the following class conditional joint prior distribution

$$p_\theta(x,z|y) = \frac{C_{\nu,m+n}}{|\Sigma_x|^{\frac{1}{2}}|\Sigma_y|^{\frac{1}{2}}}\left[1+\frac{(z-\mu_y)^\top\Sigma_y^{-1}(z-\mu_y)+(x-\mu_\theta(z))^\top\Sigma_x^{-1}(x-\mu_\theta(z))}{\nu}\right]^{-\frac{\nu+m+n}{2}},$$

with ν, n and m being the degrees of freedom of the Student's t-distribution, the dimension of the input data and the dimension of the latent space respectively. $\mu_y \in \mathbb{R}^m$ is a learnable mean vector representing class centers in latent space of dimension m. Moreover, Σ_x and Σ_y are the covariance matrices of the prior distributions over the latent and output variables.

From this joint distribution, we can derive the conditional latent prior $p(z|y)$ and decoder distribution $p_\theta(x|z,y)$ (See Appendix A)

$$p(z|y) = t_m(z|\mu_y,\Sigma_y,\nu), \quad p_\theta(x|z,y) = t_n\left(x\middle|\mu_\theta(z), \frac{\left(1+\nu^{-1}(z-\mu_y)^\top\Sigma_y^{-1}(z-\mu_y)\right)}{(1+\nu^{-1}m)}\Sigma_x, \nu+m\right).$$

By defining class-specific priors, the model ensures that each class occupies a proportionally equal latent region, countering the frequency-aligned bias of a global prior. Furthermore, as in t^3VAE, we define the posterior $q_\phi(z|x)$ as a multivariate Student's t-distribution capturing heavy-tailed structure in the latent space :

$$q_\phi(z|x) = t_m\left(z\middle|\mu_\phi(x), \frac{\Sigma_\phi(x)}{1+\nu^{-1}n}, \nu+n\right).$$

Although the posterior is defined as $q_\phi(z|x)$ without explicit class conditioning, the class-specific prior and objective derived in the following section enforce class-discriminative latent encodings.

4.2 Objective function

Harnessing Eq. (5) and the defined prior and posterior distributions of the proposed C-t^3VAE, we derive in Appendix B the following class-wise objective

$$\mathcal{L}(\gamma,y) = \mathbb{E}_x\Bigg[\mathbb{E}_z\Big[(x-\mu_\theta(z))^\top\Sigma_x^{-1}(x-\mu_\theta(z))\Big] + (\mu_\phi(x)-\mu_y)^\top\Sigma_y^{-1}(\mu_\phi(x)-\mu_y) + \frac{\nu\,\mathrm{Tr}\left(\Sigma_y^{-1}\Sigma_\phi(x)\right)}{\nu+n-2} - \frac{\nu C_1}{C_2}|\Sigma_\phi(x)|^{-\frac{\gamma}{2(1+\gamma)}}\Bigg],$$

with $C_1 = \left(C_{\nu+n,m}^\gamma\left(1+\frac{n}{\nu}\right)^{\frac{\gamma m}{2}}\frac{\nu+n+m-2}{\nu+n-2}\right)^{\frac{1}{1+\gamma}}$ and $C_2 = \left(\frac{C_{\nu,m+n}^\gamma}{|\Sigma_x|^{\frac{\gamma}{2}}|\Sigma_y|^{\frac{2\gamma+1}{2}}}\left(1+\frac{m+n}{\nu-2}\right)^{-\gamma}\right)^{\frac{1}{1+\gamma}}$.

By taking $\Sigma_x = \sigma^2 I$ and $\Sigma_y = I$, $\mathcal{L}(\gamma,y)$ objective function simplifies to :

$$\mathcal{L}(\gamma,y) = \mathbb{E}_x\left[\frac{\mathbb{E}_z\left[\|x-\mu_\theta(z)\|^2\right]}{\sigma^2} + \|\mu_\phi(x)-\mu_y\|^2 + \frac{\nu\,\mathrm{Tr}\left(\Sigma_\phi(x)\right)}{\nu+n-2} - \frac{\nu C_1}{C_2}|\Sigma_\phi(x)|^{-\frac{\gamma}{2(1+\gamma)}}\right]. \tag{9}$$

From Eq. (9), we observe that the first term aims to minimize the input reconstruction error, ensuring data fidelity, while the second term promotes alignment between the input class mean and the inferred posterior mean. The third and fourth terms jointly regularize the posterior covariance $\Sigma_\phi(x)$; specifically, the trace penalty restricts the total variance to encourage compactness, whereas the determinant term acts as a barrier function to prevent the covariance matrix from collapsing to a singularity.

Finally, we express the final loss function $\mathcal{L}(\gamma)$ over the whole dataset as :

$$\mathcal{L}(\gamma) = \sum_y \mathcal{L}(\gamma,y).$$

As in Eq. (3), here too we use uniform class weighting so the loss does not scale with empirical frequency and the generative step targets all classes uniformly.

4.3 Sampling distribution

To sample from the latent space of the C-t^3VAE, we define the following sampling distribution :

$$p^\star_\nu(z) = \sum_{y=1}^{K} \alpha_y \cdot p^\star_{\nu,y}(z) = \sum_{y=1}^{K} \alpha_y \cdot t_m(\mu_y, \tau^2 I, \nu + n), \quad \forall y, \;\; \alpha_y = \frac{1}{K}. \tag{10}$$

For the the variance parameter τ^2 in $t_m(\mu_y, \tau^2 I, \nu + n)$, we derive it theoretically. In this derivation we first derive the γ-power divergence between $p^\star_{\nu,y}(z) = t_m(\mu_y, \tau^2 I, \nu + n)$ and $q_\phi(z|x)$ for every y. Then, we compare the obtained result to the corresponding regularization terms in $\mathcal{L}(\gamma, y)$ Eq. (9) (See Appendix C for details). The obtained form is similar to the form expressed in Eq. (8).

The mixture-based sampling distribution we define in Eq. (10) with equal α_y is not a post-hoc balancing heuristic. Rather, it is the generative counterpart of promoting uniform prior mass across class-conditioned latent components. Sampling proportionally to empirical frequency would partially reintroduce frequency-aligned bias at generation time, attenuating the structural change imposed during training. Our framework also allows prioritization by modifying the mixture weights α_y when targeted generation is desired; however, in this work we focus on the balanced setting.

4.3.1 β-C-t^3VAE extension

As with t^3VAE, the class-wise objective defined in Eq. (9) can be split into a reconstruction and regularization terms. By preceding the regularization term with a β scalar, we can define a β-C-t^3VAE model thereby improving the domain of applicability of the model.

5 Experiments

This section presents quantitative and qualitative results for the C-t^3VAE model and closely related baselines (VAE, C-VAE, and t^3VAE, with their β variants). This controlled comparison isolates the contributions of our design choices:

- **VAE :** ELBO trained standard Gaussian-based VAEs.
- **C-VAE :** VAE supplemented by conditional Gaussian priors to assess the class conditioning effect without changing the prior family.
- **t^3VAE :** Student's t-distribution latent prior and γ-power divergence objective; it does not use class-conditional priors and does not allow class-conditional generation. Comparing to this model isolates the effect of conditional modeling with heavy-tailed priors.

We also evaluate β-VAE, β-C-VAE, and β-t^3VAE variants to assess the effect of tuning regularization on latent disentanglement and generative performance.

Besides, we note that we do not aim to compete with state-of-the-art diffusion or adversarial models in raw sample quality. Instead, we isolate the effect of latent prior geometry under class imbalance within a controlled variational framework. Comparing across fundamentally different architectural families would confound inductive bias, training stability, and objectives with the geometric contribution we study. Hence, our evaluation is restricted to VAE-family models and we do not benchmark against diffusion-based generators.

In the following, we first present the evaluation metrics and protocol, then analyze the latent sampling standard deviation τ used in t^3VAE and C-t^3VAE by varying τ to assess alignment between empirical and theoretical values. We then report FID comparisons across baselines with optimized hyper-parameters (the tuning of β, ν, and τ is reported in Appendix E), followed by per-class generative evaluation.

5.1 Evaluation metrics

In this section, we present the evaluation metrics used throughout this work. We report FID and generative Precision, Recall and F1 because they are the most commonly used metrics in the VAE and long-tailed generation literature, enabling direct comparison with prior work. Also, these metrics allow us to assess whether the class-conditional heavy-tailed prior in C-t^3VAE improves minority-class generation relative to baselines.

5.1.1 Fréchet Inception Distance

The Fréchet Inception Distance (FID) (Heusel et al., 2017) is a standard metric for evaluating the quality of synthetic images. It measures the similarity between the distributions of real and generated images, with a score of 0 indicating identical distributions.

To compute the FID, feature encodings of real and generated images are extracted using the InceptionV3 network (Szegedy et al., 2016), excluding the classification layer. Assuming these encodings follow a multivariate Gaussian distribution, the mean vectors μ_r and μ_s and covariance matrices Σ_r and Σ_s are estimated for the real and synthetic image sets, respectively. The FID is then calculated as:

$$FID = \|\mu_r - \mu_s\|_2^2 + \mathrm{Tr}\left(\Sigma_r + \Sigma_s - 2 \cdot (\Sigma_r \Sigma_s)^{\frac{1}{2}}\right).$$

5.1.2 Generative Precision, Recall, and F1

While FID provides a holistic measure of distributional similarity, it does not explicitly assess the quality and diversity of generated samples. To address this, generative Precision and Recall metrics have been proposed (Kynkäänniemi et al., 2019). These metrics evaluate respectively the sharpness and mode coverage of generated samples relative to the target distribution. Their harmonic mean, the F1 score, offers a balanced assessment.

Given sets of real and generated samples X_r and X_g, feature representations are extracted using a classifier network, yielding vectors δ_r and δ_g. The complete sets of feature vectors are denoted Δ_r and Δ_g, with $\Delta \in \{\Delta_r, \Delta_g\}$. A binary function is defined as:

$$f(\delta, \Delta) = \begin{cases} 1 & \text{if } \|\delta - \delta'\|_2 \leq \|\delta' - \mathrm{NN}_k(\delta', \Delta)\|_2 \text{ for at least one } \delta' \in \Delta, \\ 0 & \text{otherwise,} \end{cases}$$

where $\mathrm{NN}_k(\delta', \Delta)$ denotes the k-th nearest neighbor of δ' in Δ (we use $k = 3$ in all experiments). The function $f(\delta_g, \Delta_r)$ determines whether a generated sample appears realistic, while $f(\delta_r, \Delta_g)$ assesses whether a real sample could be reproduced by the generator. Precision and Recall are defined as:

$$\mathrm{Precision}(\Delta_r, \Delta_g) = \frac{1}{|\Delta_g|} \sum_{\delta_g \in \Delta_g} f(\delta_g, \Delta_r),$$

$$\mathrm{Recall}(\Delta_r, \Delta_g) = \frac{1}{|\Delta_r|} \sum_{\delta_r \in \Delta_r} f(\delta_r, \Delta_g).$$

Finally, the F1 score is computed as:

$$F1 = 2 \cdot \frac{\mathrm{Precision} \cdot \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}}.$$

We note that Precision, Recall, and F1 range from 0 to 1, with 1 indicating optimal performance.

5.2 Evaluation procedure

We train all models on imbalanced datasets, where the class frequency decreases as a function of the class index, and evaluate them on balanced test sets. This protocol measures robustness by assessing a model's ability to generate high-quality samples across all classes, regardless of their frequency during training.

To evaluate performance across varying complexities, we utilize SVHN (Netzer et al., 2011), CIFAR100 (Krizhevsky, 2009; Cao et al., 2019), and CelebA (Liu et al., 2015) (see Appendix D.1). SVHN serves as a tractable baseline—simple enough to ensure convergence, yet rich enough to highlight generative discrepancies. CIFAR100 provides a more challenging benchmark with high semantic diversity, particularly testing performance in low-data regimes. Finally, CelebA enables a detailed analysis of generative quality and class balance in the presence of natural attributes.

We construct Long-Tail (LT) variants of SVHN and CIFAR100 to explicitly control the training imbalance. Specifically, we induce imbalance via an exponential decay in sample counts after equalizing the initial class sizes. The imbalance ratio ρ defines the disparity between the most and least frequent classes, with the sample count M_{y_i} for class y_i given by:

$$M_{y_i} = M \cdot \rho^{-\frac{y_i - 1}{K-1}},$$

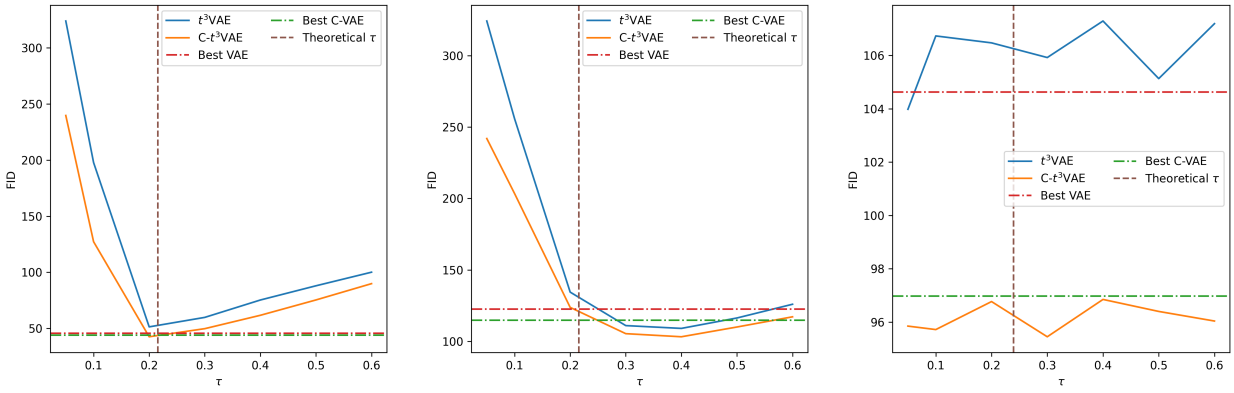
where M denotes the original sample count per class and $y_i \in \{1, \dots, K\}$ represents the class index.

For CelebA, we exploit the inherent imbalance of facial attributes. We select Mustache and Young to represent strong ($\rho \approx 25$) and moderate ($\rho \approx 3.5$) imbalance, respectively, alongside the balanced Male and Smiling attributes. We deliberately refrain from defining classes via attribute composition; such an approach would lead to a combinatorial expansion of the label space and result in prohibitive sample sparsity per class.

We employ the FID (Heusel et al., 2017) as the primary global metric to quantify the distribution shift between generated and real samples. However, as FID can be biased in low-sample regimes, we complement it with Precision, Recall, and F1-score to provide a granular, per-class evaluation of the proposed model. We detail our experimental settings in Appendix D.

5.3 τ parameter study

Figure 1 illustrates the impact of the sampling standard deviation τ on FID. Models based on the Student's t-distribution benefit from higher standard deviation on CIFAR100-LT compared to SVHN-LT. Notably, C-t^3VAE outperforms C-VAE for $\tau \in [0.25; 0.55]$ on CIFAR100-LT, $\tau \in [0.19; 0.28]$ on SVHN-LT, and for all τ values on CelebA. Additionally, it surpasses the t^3VAE FID for all τ values and across all datasets, underlining the importance of equal per-class prior mass.



(a) SVHN-LT (b) CIFAR100-LT (c) CelebA

Figure 1: FID score as a function of τ for the t^3VAE and C-t^3VAE models. Results are for the imbalance ratio $\rho = 100$ for the SVHN-LT and CIFAR100-LT, and the Mustache attribute ($\rho = 25$) for CelebA. Other imbalance ratios' results paint a similar picture and are provided in Appendix (E.3). The horizontal dashed lines is the FID value of the best performing VAE and C-VAE on each dataset and the vertical dashed line is the value of τ as derived in Eq. (8). We note that the used models in these figures have optimized β and ν hyper-parameters.

Moreover, for both Student's t-distribution based models, the optimal FID score for SVHN-LT occurs near the theoretically derived τ value. However, for the more complex CIFAR100-LT dataset, the optimal τ is higher than the theoretical value $\tau = 0.4$. The analytically derived τ provides a principled initialization grounded in divergence geometry. Dataset-specific deviations reflect encoder–decoder capacity limits rather than theoretical inconsistency, highlighting interactions between prior heaviness and representation complexity. For CelebA, τ has minimal impact on performance, likely due to the lower variability in the dataset's images (All faces are visually similar, reducing the impact of latent scale). Developing tighter theoretical characterizations of τ that account for dataset complexity and model capacity remains an interesting direction for future work.

5.4 Optimized Model Results Discussion

After optimizing the hyperparameters of the various models tested in this work, we present their generation FID scores in Table 1. We provide results for optimized β models and non-β models ($\beta = 1$) to underscore the importance of this parameter, which was not explored in the original t^3VAE work (Kim et al., 2024).

Table 1: Generation FID results on the SVHN-LT, CIFAR100-LT and CelebA datasets. For the SVHN-LT and CIFAR100-LT datasets we use different imbalance ratios $\rho \in \{100, 50, 10, 1\}$. However, for the CelebA dataset we use the Mustache, Young, Male and Smiling attributes which have imbalance ratios of 25, 3.5, 1.4 and 1 respectively. The β models undertook an optimization of the β hyper-parameter while non-β models have $\beta = 1$. All models have optimized ν and τ hyper-parameters. The attributes for the CelebA dataset column indicate which attribute is used to condition the conditional models and balance the test set.

	SVHN-LT				CIFAR100-LT				CelebA			
Models	ρ =100	50	10	1	ρ =100	50	10	1	Mustache	Young	Male	Smiling
VAE	93.89	91.91	91.66	92.16	163.66	162.91	165.47	166.46	110.58	92.01	110.58	82.05
β-VAE	47.11	49.81	45.70	43.48	122.62	123.07	123.72	124.43	104.63	92.87	87.96	83.15
C-VAE	74.75	70.40	72.30	74.16	157.90	163.67	162.09	163.24	96.98	89.17	86.17	**78.35**
β-C-VAE	48.39	46.39	43.97	**43.87**	114.88	118.89	114.89	118.21	98.35	85.53	**79.76**	78.46
t^3VAE	57.07	54.30	52.10	51.52	136.63	137.24	138.92	135.23	105.80	88.07	83.62	78.90
β-t^3VAE	51.62	49.55	48.93	45.37	109.11	107.93	108.97	**111.00**	105.86	88.21	83.83	78.89
C-t^3VAE	47.09	46.29	47.43	51.32	125.48	127.96	130.28	129.40	101.18	87.07	81.92	80.97
β-C-t^3VAE	**44.02**	**42.60**	**42.01**	44.49	**103.25**	**102.99**	**105.92**	112.37	**95.82**	**82.61**	81.65	80.08

From Table 1, t^3VAE improves over VAE, highlighting the generative advantage of the Student's t-distribution prior over a Gaussian one, in addition to the reconstruction advantage noted in (Kim et al., 2024). Optimizing β improves t^3VAE FID over VAE on CIFAR100-LT and CelebA, while remaining competitive on SVHN-LT. Qualitative results in Figure 2 show that t^3VAE produces sharper synthetic images on CelebA than VAE.

Figure 2: Sample synthetic images from the optimized VAE and t^3VAE models trained on the CelebA dataset. No class conditioning is possible for these models.

For class-conditional models, optimized C-t^3VAE yields strong FID improvements across imbalanced settings on SVHN-LT and CIFAR100-LT, and on heavily imbalanced CelebA attributes. As shown in Table 1, it achieves gains of up to 4, 5, and 10 FID points over β-t^3VAE on imbalanced settings of SVHN-LT, CIFAR100-LT, and CelebA, respectively. This supports the role of uniform prior mass allocation in high-imbalance regimes. Moreover, β-C-t^3VAE reduces FID by up to 4 and 15 points over C-VAE on SVHN-LT and CIFAR100-LT, respectively. For CelebA, β-C-t^3VAE achieves the best results on heavily imbalanced

attributes like Mustache, indicating improved generation for underrepresented classes. The gain over C-VAE follows from the Student's t-distribution latent prior and its ability to better capture intra-class long-tail structure. Qualitative samples of conditional optimized models (Figure 3) show sharper facial features for C-t^3VAE than C-VAE, notably on the imbalanced Mustache attribute. Overall, C-t^3VAE exhibits the most consistent performance in high-imbalance regimes within the VAE family.

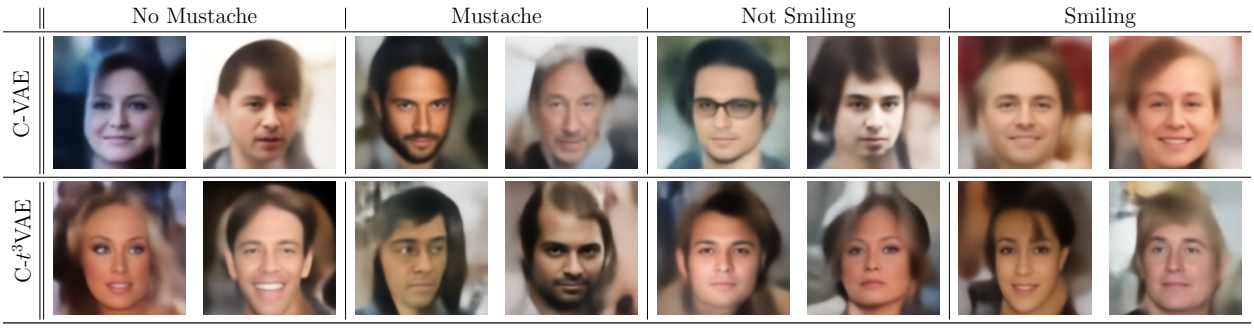
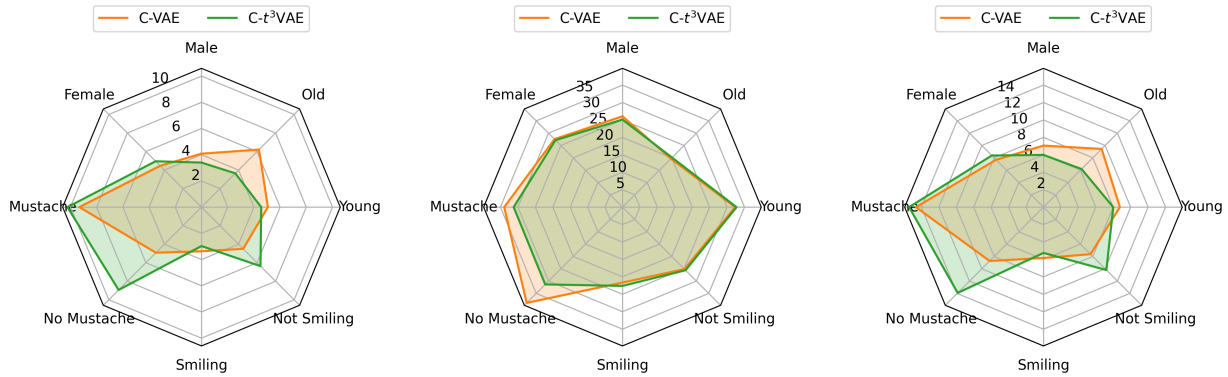


Figure 3: Sample synthetic images for the optimized C-VAE and C-t^3VAE trained on specific attributes of the CelebA dataset.

5.5 Per-class evaluation

In this section, we evaluate the conditional models on a per-class basis. Since FID can be biased on small datasets and offers limited insight as a single scalar metric, we rely on Precision, Recall, and F1 metrics (Kynkäänniemi et al., 2019). Our results on CelebA are shown in Figure 4, with additional results for SVHN-LT and CIFAR100-LT included in Appendix F.



(a) Recall (b) Precision (c) F1 score

Figure 4: Per-class generative metrics on CelebA after optimization of all hyper-parameters notably β, ν and τ. We note that the imbalance ratio of the Mustache, Young, Male and Smiling factors ρ are 25, 3.5, 1.4 and 1 respectively.

On CelebA (Figure 4), C-t^3VAE improves Recall and F1 on the most imbalanced attribute (Mustache), but not on more balanced ones (Male, Smiling), which is expected. When class imbalance is moderate (e.g., Young attribute with $\rho = 3.5$), C-VAE slightly outperforms C-t^3VAE, indicating that Gaussian priors suffice when encoder variance and decoder flexibility can compensate for small frequency skew. This suggests the presence of a regime where Gaussian priors may suffice. To explore this, we vary the imbalance ratio on SVHN-LT from 100 to 1 and plot the results in Figure 5. This figure shows a threshold around $\rho \approx 5$: below it C-VAE can be better, while above it C-t^3VAE has the advantage. As the imbalance ratio increases, the FID gap grows in favor of C-t^3VAE. For moderate imbalance ($\rho < 5$), encoder variance and decoder flexibility can compensate for

frequency skew; beyond this regime, posterior–prior mismatch grows nonlinearly, and Gaussian priors increasingly compress minority-class latent regions. Heavy-tailed class-conditional priors mitigate this compression by permitting larger dispersion without penalizing minority likelihood, explaining the observed transition.

For SVHN-LT (Figures in Appendix F), C-t^3VAE achieves higher Recall and competitive Precision compared to C-VAE, indicating better mode coverage while maintaining image quality, which yields higher F1 scores, especially on tail classes. For CIFAR100-LT (Figures in Appendix F), C-VAE often attains high Precision but near-zero Recall, indicating mode collapse. In contrast, C-t^3VAE preserves Recall at the cost of slightly lower Precision, resulting in improved F1 scores. Consequently, on SVHN-LT, CIFAR100-LT, and CelebA, C-t^3VAE is the most reliable method in our study for high-imbalance settings within the VAE family.

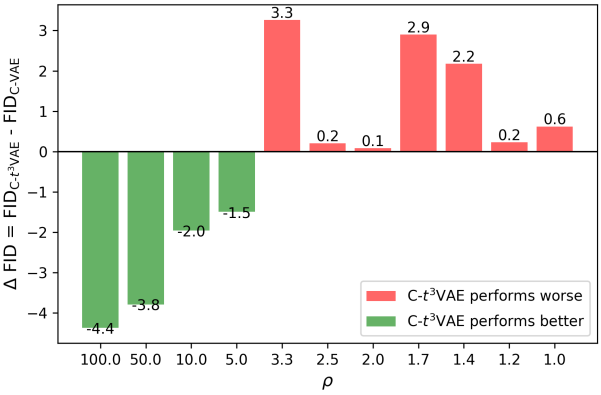


Figure 5: C-t^3VAE versus C-VAE models under varying imbalance ratios on the SVHN-LT dataset.

6 Conclusion

In this work, we introduced C-t^3VAE. This class-conditional generative model mitigates frequency-aligned latent allocation effects observed under imbalance. It uses per-class Student's t-distributions in the latent space, paired with a theoretically derived, equal-weight sampling scheme. This design allocates uniform prior mass across all classes while capturing heavy-tailed intra-class variations. Furthermore, these structural improvements do not sacrifice efficiency. The computational complexity of C-t^3VAE matches that of standard conditional VAEs.

We evaluated our model on SVHN-LT, CIFAR100-LT, and CelebA. After optimizing the hyperparameters β, ν, and τ, C-t^3VAE consistently outperforms baseline models in high-imbalance regimes. It achieves up to a 15-point FID improvement over t^3VAE and conditional VAEs. Additionally, per-class Precision, Recall, and F1 metrics confirm superior mode coverage for tail classes. We also identified a critical imbalance ratio threshold at $\rho \approx 5$. For milder imbalances ($\rho \lesssim 5$), Gaussian-based models remain competitive. However, as imbalance grows beyond this threshold, C-t^3VAE demonstrates a significant and widening advantage.

Future research should explore extending the C-t^3VAE framework to multi-label settings. The MultiFacet VAE model (Falck et al., 2021) provides a promising starting point for handling complex, overlapping annotations. Furthermore, our experiments show that the sampling variance τ heavily impacts image quality. This parameter currently requires dataset-specific tuning. Therefore, developing tighter theoretical bounds and adaptive latent-space sampling methods remains a crucial next step for high-fidelity VAE generation.

References

Najmeh Abiri and Mattias Ohlsson. Variational auto-encoders with student's t-prior. *arXiv preprint arXiv:2004.02581*, 2020.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International Conference on Machine Learning*, 2017.

Siddarth Asokan and Chandra Seelamantula. Teaching a GAN what not to learn. *Advances in Neural Information Processing Systems*, 2020.

Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. In *Advances in Neural Information Processing Systems*, 2019.

Clément Chadebec, Elina Thibeau-Sutre, Ninon Burgos, and Stéphanie Allassonnière. Data augmentation in high dimensional low sample size setting using a geometry-based variational autoencoder. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3):2879–2896, 2023.

Tim R. Davidson, Luca Falorsi, Nicola De Cao, Thomas Kipf, and Jakub M. Tomczak. Hyperspherical variational auto-encoders. *Uncertainty in Artificial Intelligence Conference*, 2018.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 2021.

Nat Dilokthanakul, Pedro AM Mediano, Marta Garnelo, Matthew CH Lee, Hugh Salimbeni, Kai Arulkumaran, and Murray Shanahan. Deep unsupervised clustering with gaussian mixture variational autoencoders. *arXiv preprint arXiv:1611.02648*, 2016.

Shinto Eguchi. Pythagoras theorem in information geometry and applications to generalized linear models. In *Information Geometry*, volume 45 of *Handbook of Statistics*, pp. 15–42. Elsevier, 2021.

Fabian Falck, Haoting Zhang, Matthew Willetts, George Nicholson, Christopher Yau, and Chris C Holmes. Multi-facet clustering variational autoencoders. *Advances in Neural Information Processing Systems*, 2021.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, 2014.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. In *Advances in Neural Information Processing Systems*, 2017.

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.

Priyank Jaini, Ivan Kobyzev, Yaoliang Yu, and Marcus Brubaker. Tails of lipschitz triangular flows. In *International Conference on Machine Learning*, 2020.

Juno Kim, Jaehyuk Kwon, Mincheol Cho, Hyunjong Lee, and Joong-Ho Won. t^3-variational autoencoder: Learning heavy-tailed data with student's t and power divergence. In *International Conference on Learning Representations*, 2024.

Diederik P Kingma and Max Welling. Auto-Encoding Variational Bayes, 2013. arXiv:1312.6114.

Durk P Kingma, Shakir Mohamed, Danilo Jimenez Rezende, and Max Welling. Semi-supervised learning with deep generative models. *Advances in Neural Information Processing Systems*, 2014.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, 2009.

Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved precision and recall metric for assessing generative models. *Advances in Neural Information Processing Systems*, 2019.

Zinan Lin, Ashish Khetan, Giulia Fanti, and Sewoong Oh. PacGAN: The power of two samples in generative adversarial networks. *Advances in Neural Information Processing Systems*, 2018.

Zinan Lin, Hao Liang, Giulia Fanti, and Vyas Sekar. RareGAN: Generating samples for rare classes. In *AAAI Conference on Artificial Intelligence*, 2022.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep Learning Face Attributes in the Wild. In *International Conference on Computer Vision*, December 2015.

Alireza Makhzani, Jonathon Shlens, Navdeep Jaitly, Ian Goodfellow, and Brendan Frey. Adversarial autoencoders. *International Conference on Learning Representations workshop*, 2016.

Emile Mathieu, Tom Rainforth, Nana Siddharth, and Yee Whye Teh. Disentangling disentanglement in variational autoencoders. In *International Conference on Machine Learning*, 2019a.

Emile Mathieu, Tom Rainforth, Nana Siddharth, and Yee Whye Teh. Disentangling disentanglement in variational autoencoders. In *International Conference on Machine Learning*, 2019b.

Dwij Mehta, Aditya Mehta, and Pratik Narang. Ldfacenet: Latent diffusion-based network for high-fidelity deepfake generation. In *International Conference on Pattern Recognition*. Springer, 2024.

Ranjita Naik and Besmira Nushi. Social biases through the text-to-image generation lens. In *AAAI/ACM Conference on AI, Ethics, and Society*, 2023.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y. Ng. Reading digits in natural images with unsupervised feature learning. In *Advances in Neural Information Processing Systems Workshop on Deep Learning and Unsupervised Feature Learning*, 2011.

Kushagra Pandey, Jaideep Pathak, Yilun Xu, Stephan Mandt, Mike Pritchard, Arash Vahdat, and Morteza Mardani. Heavy-tailed diffusion models. In *International Conference on Learning Representations*, 2025.

Walter HL Pinaya, Petru-Daniel Tudosiu, Jessica Dafflon, Pedro F Da Costa, Virginia Fernandez, Parashkev Nachev, Sebastien Ourselin, and M Jorge Cardoso. Brain imaging generation with latent diffusion models. In *MICCAI Workshop on Deep Generative Models*. Springer, 2022.

Yiming Qin, Huangjie Zheng, Jiangchao Yao, Mingyuan Zhou, and Ya Zhang. Class-balancing diffusion models. In *Computer Vision and Pattern Recognition Conference*, 2023.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Computer Vision and Pattern Recognition Conference*, 2022.

Amrutha Saseendran, Kathrin Skubch, Stefan Falkner, and Margret Keuper. Shape your space: A gaussian mixture regularization approach to deterministic autoencoders. In *Advances in Neural Information Processing Systems*, 2021.

Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. In *Advances in Neural Information Processing Systems*, 2015.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Computer Vision and Pattern Recognition Conference*, June 2016.

Hiroshi Takahashi, Tomoharu Iwata, Yuki Yamanaka, Masanori Yamada, and Satoshi Yagi. Student-t variational autoencoder for robust density estimation. In *International Joint Conference on Artificial Intelligence*, 2018.

Hiroshi Takahashi, Tomoharu Iwata, Yuki Yamanaka, Masanori Yamada, and Satoshi Yagi. Variational autoencoder with implicit optimal priors. In *AAAI Conference on Artificial Intelligence*, 2019.

Edric Tam and David B Dunson. On the statistical capacity of deep generative models. *arXiv preprint arXiv:2501.07763*, 2025.

Arash Vahdat and Jan Kautz. Nvae: A deep hierarchical variational autoencoder. *Advances in Neural Information Processing Systems*, 2020.

Aaron van den Oord, Oriol Vinyals, and koray kavukcuoglu. Neural discrete representation learning. In *Advances in Neural Information Processing Systems*, 2017.

Tianjiao Zhang, Huangjie Zheng, Jiangchao Yao, Xiangfeng Wang, Mingyuan Zhou, Ya Zhang, and Yanfeng Wang. Long-tailed diffusion models with oriented calibration. In *International Conference on Learning Representations*, 2024.

Supplementary Material

A Priors derivations

In this section, we present our derivations of the different prior distributions defining our proposed C-t^3-VAE model. Starting from the proposed joint distribution :

$$p_\theta(x,z|y) = \frac{C_{\nu,m+n}}{|\Sigma_x|^{\frac{1}{2}}|\Sigma_y|^{\frac{1}{2}}}\left[1+\frac{(z-\mu_y)^\top\Sigma_y^{-1}(z-\mu_y)+(x-\mu_\theta(z))^\top\Sigma_x^{-1}(x-\mu_\theta(z))}{\nu}\right]^{-\frac{\nu+m+n}{2}}.$$

To calculate the prior distribution on the latent space we marginalize out x as follows :

$$\begin{aligned}
p(z|y) &= \int p_\theta(x,z|y)dx \\
&= \int C_{\nu,m+n}|\Sigma_x|^{-\frac{1}{2}}|\Sigma_y|^{-\frac{1}{2}}\left[1+\frac{(z-\mu_y)^\top\Sigma_y^{-1}(z-\mu_y)}{\nu}+\frac{(x-\mu_\theta(z))^\top\Sigma_x^{-1}(x-\mu_\theta(z))}{\nu}\right]^{-\frac{\nu+m+n}{2}}dx \\
&= C_{\nu,m+n}|\Sigma_x|^{-\frac{1}{2}}|\Sigma_y|^{-\frac{1}{2}}\left[1+\frac{1}{\nu}(z-\mu_y)^\top\Sigma_y^{-1}(z-\mu_y)\right]^{-\frac{\nu+m+n}{2}} \\
&\quad\times\int\left(1+\frac{(1+\nu^{-1}m)(x-\mu_\theta(z))^\top\Sigma_x^{-1}(x-\mu_\theta(z))}{\left(1+\nu^{-1}(z-\mu_y)^\top\Sigma_y^{-1}(z-\mu_y)\right)(\nu+m)}\right)^{-\frac{\nu+m+n}{2}}dx.
\end{aligned}$$

Given that :

$$\begin{aligned}
&\int C_{\nu+m,n}|\Sigma|^{-\frac{1}{2}}\left(1+\frac{(x-\mu)^\top\Sigma^{-1}(x-\mu)}{\nu+m}\right)^{-\frac{\nu+m+n}{2}}dx = 1 \\
&\Rightarrow \int\left(1+\frac{(x-\mu)^\top\Sigma^{-1}(x-\mu)}{\nu+m}\right)^{-\frac{\nu+m+n}{2}}dx = C^{-1}_{\nu+m,n}|\Sigma|^{\frac{1}{2}},
\end{aligned}$$

and when setting :

$$\Sigma^{-1} = \frac{(1+\nu^{-1}m)\Sigma_x^{-1}}{1+\nu^{-1}(z-\mu_y)^\top\Sigma_y^{-1}(z-\mu_y)},$$

We get :

$$\begin{aligned}
&\int\left(1+\frac{(1+\nu^{-1}m)(x-\mu_\theta(z))^\top\Sigma_x^{-1}(x-\mu_\theta(z))}{\left(1+\nu^{-1}(z-\mu_y)^\top\Sigma_y^{-1}(z-\mu_y)\right)(\nu+m)}\right)^{-\frac{\nu+m+n}{2}}dx \\
&\qquad = C^{-1}_{\nu+m,n}\left|\left(\frac{(1+\nu^{-1}m)\Sigma_x^{-1}}{1+\nu^{-1}(z-\mu_y)^\top\Sigma_y^{-1}(z-\mu_y)}\right)^{-1}\right|^{\frac{1}{2}} \\
&\qquad = C^{-1}_{\nu+m,n}\left|\frac{1+\nu^{-1}(z-\mu_y)^\top\Sigma_y^{-1}(z-\mu_y)}{(1+\nu^{-1}m)}\Sigma_x\right|^{\frac{1}{2}} \\
&\qquad = C^{-1}_{\nu+m,n}\left(\frac{1+\nu^{-1}(z-\mu_y)^\top\Sigma_y^{-1}(z-\mu_y)}{1+\nu^{-1}m}\right)^{\frac{n}{2}}|\Sigma_x|^{\frac{1}{2}}.
\end{aligned}$$

Therefore, $p(z|y)$ simplifies to :

$$\begin{aligned}
p(z|y) &= C_{\nu,m+n}|\Sigma_x|^{-\frac{1}{2}}|\Sigma_y|^{-\frac{1}{2}}\left[1+\frac{1}{\nu}(z-\mu_y)^\top\Sigma_y^{-1}(z-\mu_y)\right]^{-\frac{\nu+m+n}{2}} C_{\nu+m,n}^{-1}\,|\Sigma_x|^{\frac{1}{2}} \\
&\quad\times\left(\frac{1+\nu^{-1}(z-\mu_y)^\top\Sigma_y^{-1}(z-\mu_y)}{1+\nu^{-1}m}\right)^{\frac{n}{2}} \\
&= C_{\nu,m+n}C_{\nu+m,n}^{-1}\left(1+\frac{m}{\nu}\right)^{-\frac{n}{2}}|\Sigma_y|^{-\frac{1}{2}}\left[1+\frac{1}{\nu}(z-\mu_y)^\top\Sigma_y^{-1}(z-\mu_y)\right]^{-\frac{\nu+m}{2}} \\
&= C_{\nu,m}|\Sigma_y|^{-\frac{1}{2}}\left[1+\frac{1}{\nu}(z-\mu_y)^\top\Sigma_y^{-1}(z-\mu_y)\right]^{-\frac{\nu+m}{2}} \\
&= t_m(z|\mu_y,\Sigma_y,\nu).
\end{aligned}$$

Here and in the following, we use the fact

$$C_{\nu,m+n} = C_{\nu+m,n}C_{\nu,m}\left(1+\frac{m}{\nu}\right)^{\frac{n}{2}}.$$

Besides, the prior distribution over the output of the decoder model $p(x|z,y)$ can be derived as follows :

$$\begin{aligned}
p_\theta(x|z,y) &= \frac{p_\theta(x,z|y)}{p(z|y)} \\
&= \frac{C_{\nu,m+n}}{|\Sigma_x|^{\frac{1}{2}}|\Sigma_y|^{\frac{1}{2}}}\left[1+\frac{(z-\mu_y)^\top\Sigma_y^{-1}(z-\mu_y)+(x-\mu_\theta(z))^\top\Sigma_x^{-1}(x-\mu_\theta(z))}{\nu}\right]^{-\frac{\nu+m+n}{2}} \\
&\quad\times C_{\nu,m}^{-1}|\Sigma_y|^{\frac{1}{2}}\left[1+\frac{1}{\nu}(z-\mu_y)^\top\Sigma_y^{-1}(z-\mu_y)\right]^{\frac{\nu+m}{2}} \\
&= C_{\nu+m,n}|\Sigma_x|^{-\frac{1}{2}}\left(1+\frac{m}{\nu}\right)^{\frac{n}{2}}\left[1+\frac{1}{\nu}(z-\mu_y)^\top\Sigma_y^{-1}(z-\mu_y)\right]^{-\frac{n}{2}} \\
&\quad\times\left(1+\frac{(1+\nu^{-1}m)(x-\mu_\theta(z))^\top\Sigma_x^{-1}(x-\mu_\theta(z))}{\left(1+\nu^{-1}(z-\mu_y)^\top\Sigma_y^{-1}(z-\mu_y)\right)(\nu+m)}\right)^{-\frac{\nu+m+n}{2}} \\
&= t_n\left(x\middle|\mu_\theta(z),\frac{\left(1+\nu^{-1}(z-\mu_y)^\top\Sigma_y^{-1}(z-\mu_y)\right)}{(1+\nu^{-1}m)}\Sigma_x,\nu+m\right).
\end{aligned}$$

B Loss function derivation

In this section, we derive the loss function of C-t^3-VAE. We start by calculating the different double integrals $\iint p_\theta(x,z|y)^{1+\gamma}dxdz$, $\iint q_\phi(x,z|y)p_\theta(x,z|y)^{\gamma}dxdz$, and $\iint q_\phi(x,z|y)^{1+\gamma}dxdz$.

Firstly,

$$\begin{aligned}
\iint p_\theta(x,z|y)^{1+\gamma}dxdz &= \mathbb{E}_{z\sim p(z|y)}\mathbb{E}_{x\sim p_\theta(x|z,y)}\left[p_\theta(x,z|y)^{\gamma}\right] \\
&= C^{\gamma}_{\nu,m+n}|\Sigma_x|^{-\frac{\gamma}{2}}|\Sigma_y|^{-\frac{\gamma}{2}}\mathbb{E}_z\mathbb{E}_x\Bigg[1+\frac{(z-\mu_y)^\top\Sigma_y^{-1}(z-\mu_y)}{\nu} \\
&\quad + \frac{(x-\mu_\theta(z))^\top\Sigma_x^{-1}(x-\mu_\theta(z))}{\nu}\Bigg] \\
&= C^{\gamma}_{\nu,m+n}|\Sigma_x|^{-\frac{\gamma}{2}}|\Sigma_y|^{-\frac{\gamma}{2}}\mathbb{E}_z\Bigg[1+\nu^{-1}(z-\mu_y)^\top\Sigma_y^{-1}(z-\mu_y) \\
&\quad + \nu^{-1}\mathbb{E}_x\left[\mathrm{Tr}(\Sigma_x^{-1}(x-\mu_\theta(z))(x-\mu_\theta(z))^\top)\right]\Bigg] \\
&= C^{\gamma}_{\nu,m+n}|\Sigma_x|^{-\frac{\gamma}{2}}|\Sigma_y|^{-\frac{\gamma}{2}}\mathbb{E}_z\Bigg[1+\nu^{-1}(z-\mu_y)^\top\Sigma_y^{-1}(z-\mu_y) \\
&\quad + \nu^{-1}\,\mathrm{Tr}\left(\Sigma_x^{-1}\Sigma_x\frac{\nu+m}{\nu+m-2}\frac{\left(1+\nu^{-1}(z-\mu_y)^\top\Sigma_y^{-1}(z-\mu_y)\right)}{(1+\nu^{-1}m)}\right)\Bigg]
\end{aligned}$$

Here, we use the following identities

$$(k-p)^\top H^{-1}(k-p) = \mathrm{Tr}\left(H^{-1}(k-p)(k-p)^\top\right); \qquad \mathbb{E}[\mathrm{Tr}(\cdot)] = \mathrm{Tr}(\mathbb{E}[\cdot])$$

and the covariance of a multivariate Student's t distribution $p\sim t(\mu;\Sigma;\nu)$ is $\frac{\nu}{\nu-2}\Sigma$. Consequently, and after a few simplifications we get

$$\begin{aligned}
\iint p_\theta(x,z|y)^{1+\gamma}dxdz &= C^{\gamma}_{\nu,m+n}|\Sigma_x|^{-\frac{\gamma}{2}}|\Sigma_y|^{-\frac{\gamma}{2}}\mathbb{E}_z\Bigg[1+\nu^{-1}(z-\mu_y)^\top\Sigma_y^{-1}(z-\mu_y) \\
&\quad + \frac{n}{\nu+m-2}\left(1+\nu^{-1}(z-\mu_y)^\top\Sigma_y^{-1}(z-\mu_y)\right)\Bigg] \\
&= C^{\gamma}_{\nu,m+n}|\Sigma_x|^{-\frac{\gamma}{2}}|\Sigma_y|^{-\frac{\gamma}{2}}\mathbb{E}_z\Bigg[\left(1+\frac{n}{\nu+m-2}\right) \\
&\quad \times\left(1+\nu^{-1}(z-\mu_y)^\top\Sigma_y^{-1}(z-\mu_y)\right)\Bigg] \\
&= C^{\gamma}_{\nu,m+n}|\Sigma_x|^{-\frac{\gamma}{2}}|\Sigma_y|^{-\frac{\gamma}{2}}\left(1+\frac{n}{\nu+m-2}\right) \\
&\quad \times\left(1+\nu^{-1}\mathbb{E}_z\left[(z-\mu_y)^\top\Sigma_y^{-1}(z-\mu_y)\right]\right) \\
&= C^{\gamma}_{\nu,m+n}|\Sigma_x|^{-\frac{\gamma}{2}}|\Sigma_y|^{-\frac{\gamma}{2}}\left(1+\frac{n}{\nu+m-2}\right) \\
&\quad \times\left(1+\nu^{-1}\mathbb{E}_z\left[\mathrm{Tr}(\Sigma_y^{-1}(z-\mu_y)(z-\mu_y)^\top\right]\right) \\
&= C^{\gamma}_{\nu,m+n}|\Sigma_x|^{-\frac{\gamma}{2}}|\Sigma_y|^{-\frac{\gamma}{2}}\left(1+\frac{n}{\nu+m-2}\right)\left(1+\frac{m}{\nu-2}\right) \\
&= C^{\gamma}_{\nu,m+n}|\Sigma_x|^{-\frac{\gamma}{2}}|\Sigma_y|^{-\frac{\gamma}{2}}\left(1+\frac{m+n}{\nu-2}\right).
\end{aligned}$$

Secondly,

$$\begin{aligned}\iint q_\phi(x,z|y)p_\theta(x,z|y)^\gamma dxdz &= \mathbb{E}_{x\sim p_{data}}\mathbb{E}_{z\sim q(z|x)}\left[p_\theta(x,z|y)^\gamma\right] \\ &= C^\gamma_{\nu,m+n}|\Sigma_x|^{-\frac{\gamma}{2}}|\Sigma_y|^{-\frac{\gamma}{2}}\mathbb{E}_x\mathbb{E}_z\Bigg[1+\frac{(z-\mu_y)^\top\Sigma_y^{-1}(z-\mu_y)}{\nu} \\ &\quad + \frac{(x-\mu_\theta(z))^\top\Sigma_x^{-1}(x-\mu_\theta(z))}{\nu}\Bigg] \\ &= C^\gamma_{\nu,m+n}|\Sigma_x|^{-\frac{\gamma}{2}}|\Sigma_y|^{-\frac{\gamma}{2}}\mathbb{E}_x\Bigg[1+\frac{1}{\nu}\mathbb{E}_z\left[(z-\mu_y)^\top\Sigma_y^{-1}(z-\mu_y)\right] \\ &\quad + \frac{1}{\nu}\mathbb{E}_z\left[(x-\mu_\theta(z))^\top\Sigma_x^{-1}(x-\mu_\theta(z))\right]\Bigg].\end{aligned}$$

Simplifying $\mathbb{E}_z\left[(z-\mu_y)^\top\Sigma_y^{-1}(z-\mu_y)\right]$:

$$\begin{aligned}\mathbb{E}_z\left[(z-\mu_y)^\top\Sigma_y^{-1}(z-\mu_y)\right] &= \mathbb{E}_z\left[\mathrm{Tr}\left(\Sigma_y^{-1}(z-\mu_y)(z-\mu_y)^\top\right)\right] \\ &= \mathbb{E}_z\left[\mathrm{Tr}\left(\Sigma_y^{-1}(z-\mu(x)+\mu(x)-\mu_y)(z-\mu(x)+\mu(x)-\mu_y)^\top\right)\right] \\ &= \mathbb{E}_z[\mathrm{Tr}(\Sigma_y^{-1}((z-\mu(x))(z-\mu(x))^\top + (z-\mu(x))(\mu(x)-\mu_y)^\top \\ &\quad + (\mu(x)-\mu_y)(z-\mu(x))^\top + (\mu(x)-\mu_y)(\mu(x)-\mu_y)^\top))] \\ &= \frac{\nu}{\nu+n-2}\,\mathrm{Tr}\left(\Sigma_y^{-1}\Sigma_\phi(x)\right) + (\mu(x)-\mu_y)^\top\Sigma_y^{-1}(\mu(x)-\mu_y).\end{aligned}$$

Then, $\iint q(x,z|y)p_\theta(x,z|y)^\gamma dxdz$ simplifies to :

$$\begin{aligned}\iint q_\phi(x,z|y)p_\theta(x,z|y)^\gamma dxdz &= C^\gamma_{\nu,m+n}|\Sigma_x|^{-\frac{\gamma}{2}}|\Sigma_y|^{-\frac{\gamma}{2}}\mathbb{E}_x\Bigg[1+\frac{1}{\nu}\mathbb{E}_z\left[(z-\mu_y)^\top\Sigma_y^{-1}(z-\mu_y)\right] \\ &\quad + \frac{1}{\nu}\mathbb{E}_z\left[(x-\mu_\theta(z))^\top\Sigma_x^{-1}(x-\mu_\theta(z))\right]\Bigg]\end{aligned}$$

$$\begin{aligned}\iint q_\phi(x,z|y)p_\theta(x,z|y)^\gamma dxdz = C^\gamma_{\nu,m+n}|\Sigma_x|^{-\frac{\gamma}{2}}|\Sigma_y|^{-\frac{\gamma}{2}}\,|\Sigma_y|^{-\frac{1}{2}}\,\mathbb{E}_x\Bigg[1+\frac{1}{\nu}\frac{\nu\,\mathrm{Tr}\left(\Sigma_y^{-1}\Sigma_\phi(x)\right)}{\nu+n-2} \\ + \frac{(\mu(x)-\mu_y)^\top\Sigma_y^{-1}(\mu(x)-\mu_y)}{\nu} + \frac{\mathbb{E}_z\left[(x-\mu_\theta(z))^\top\Sigma_x^{-1}(x-\mu_\theta(z))\right]}{\nu}\Bigg].\end{aligned}$$

Finally, the third term $\iint q_(x,z|y)^{1+\gamma}dxdz$ is

$$\iint q_\phi(x,z|y)^{1+\gamma}dxdz = C^\gamma_{\nu+n,m}\left(1+\frac{n}{\nu}\right)^{\frac{\gamma m}{2}}\left(1+\frac{m}{\nu+n-2}\right)\int|\Sigma_\phi(x)|^{-\frac{\gamma}{2}}p_{data}(x)^{1+\gamma}dx,$$

where this last double integral is equal to the one computed for the t^3-VAE.

Equipped with these formulas we can calculate the entropy $\mathcal{H}_\gamma$, cross-entropy $\mathcal{C}_\gamma$ and the γ-divergence $\mathcal{D}(q||p)$ of our model. Firstly,

$$\begin{aligned}\mathcal{H}_\gamma &= -\left(\iint q(x,z)^{1+\gamma}dxdz\right)^{\frac{1}{1+\gamma}} \\ &= -C^{\frac{\gamma}{1+\gamma}}_{\nu+n,m}\left(1+\frac{n}{\nu}\right)^{\frac{\gamma m}{2(1+\gamma)}}\left(1+\frac{m}{\nu+n-2}\right)^{\frac{1}{1+\gamma}}\left(\int|\Sigma_\phi(x)|^{-\frac{\gamma}{2}}p_{data}(x)^{1+\gamma}dx\right)^{\frac{1}{1+\gamma}},\end{aligned}$$

Which is similar to the one calculated in the t^3VAE model.

Secondly,

$$\begin{aligned}
\mathcal{C}_\gamma &= -\left(\iint q(x,z|y)p_\theta(x,z|y)^\gamma dxdz\right)\left(\iint p_\theta(x,z|y)^{1+\gamma}\right)^{-\frac{\gamma}{1+\gamma}} \\
&= -C^\gamma_{\nu,m+n}|\Sigma_x|^{-\frac{\gamma}{2}}|\Sigma_y|^{-\frac{\gamma}{2}-\frac{1}{2}}\mathbb{E}_x\Bigg[1+\frac{1}{\nu}\frac{\nu\,\mathrm{Tr}\left(\Sigma_y^{-1}\Sigma_\phi(x)\right)}{\nu+n-2}+\frac{(\mu(x)-\mu_y)^\top\Sigma_y^{-1}(\mu(x)-\mu_y)}{\nu} \\
&\quad+\frac{1}{\nu}\mathbb{E}_z\left[(x-\mu_\theta(z))^\top\Sigma_x^{-1}(x-\mu_\theta(z))\right]\Bigg]\left(C^\gamma_{\nu,m+n}|\Sigma_x|^{-\frac{\gamma}{2}}|\Sigma_y|^{-\frac{\gamma}{2}}\left(1+\frac{m+n}{\nu-2}\right)\right)^{-\frac{\gamma}{1+\gamma}} \\
&= -\left(C^\gamma_{\nu,m+n}|\Sigma_x|^{-\frac{\gamma}{2}}|\Sigma_y|^{-\frac{\gamma}{2}}\right)^{\frac{1}{1+\gamma}}|\Sigma_y|^{-\frac{1}{2}}\left(1+\frac{m+n}{\nu-2}\right)^{-\frac{\gamma}{1+\gamma}}\mathbb{E}_x\Bigg[1+\frac{1}{\nu}\Bigg(\frac{\nu\,\mathrm{Tr}\left(\Sigma_y^{-1}\Sigma_\phi(x)\right)}{\nu+n-2} \\
&\quad+(\mu(x)-\mu_y)^\top\Sigma_y^{-1}(\mu(x)-\mu_y)+\mathbb{E}_z\left[(x-\mu_\theta(z))^\top\Sigma_x^{-1}(x-\mu_\theta(z))\right]\Bigg)\Bigg].
\end{aligned}$$

Hence, we can define our divergence as :

$$\begin{aligned}
\mathcal{D}_\gamma(q||p) &= \frac{C_1}{\gamma}\left(\int|\Sigma_\phi(x)|^{-\frac{\gamma}{2}}p_{data}(x|y)^{1+\gamma}dx\right)^{\frac{1}{1+\gamma}}-\frac{C_2}{\gamma}\mathbb{E}_x\Bigg[1+\frac{1}{\nu}\Bigg(\frac{\nu\,\mathrm{Tr}\left(\Sigma_y^{-1}\Sigma_\phi(x)\right)}{\nu+n-2} \\
&\quad+(\mu(x)-\mu_y)(\mu(x)-\mu_y)^\top+\mathbb{E}_z\left[(x-\mu_\theta(z))^\top\Sigma_x^{-1}(x-\mu_\theta(z))\right]\Bigg)\Bigg] \\
&= \mathbb{E}_x\Bigg[\frac{C_1}{\gamma}|\Sigma_\phi(x)|^{-\frac{\gamma}{2(1+\gamma)}}-\frac{C_2}{\gamma}\Bigg(1+\frac{1}{\nu}\Bigg(\frac{\nu}{\nu+n-2}\,\mathrm{Tr}\left(\Sigma_y^{-1}\Sigma_\phi(x)\right) \\
&\quad+(\mu(x)-\mu_y)^\top\Sigma_y^{-1}(\mu(x)-\mu_y)+\mathbb{E}_z\left[(x-\mu_\theta(z))^\top\Sigma_x^{-1}(x-\mu_\theta(z))\right]\Bigg)\Bigg),\Bigg]
\end{aligned}$$

with C_1 and C_2 being :

$$\begin{aligned}
C_1 &= C^{\frac{\gamma}{1+\gamma}}_{\nu+n,m}\left(1+\frac{n}{\nu}\right)^{\frac{\gamma m}{2(1+\gamma)}}\left(1+\frac{m}{\nu+n-2}\right)^{\frac{1}{1+\gamma}} \\
C_2 &= \left(C^\gamma_{\nu,m+n}|\Sigma_x|^{-\frac{\gamma}{2}}|\Sigma_y|^{-\frac{2\gamma+1}{2}}\right)^{\frac{1}{1+\gamma}}\left(1+\frac{m+n}{\nu-2}\right)^{-\frac{\gamma}{1+\gamma}}.
\end{aligned}$$

On that account, the loss function for a class y is :

$$\begin{aligned}
\mathcal{L}(\gamma,y) &= -\frac{\nu\gamma}{C_2}\cdot\mathcal{D}_\gamma(q||p) \\
&= \mathbb{E}_x\Bigg[\mathbb{E}_z\left[(x-\mu_\theta(z))^\top\Sigma_x^{-1}(x-\mu_\theta(z))\right]+(\mu(x)-\mu_y)^\top\Sigma_y^{-1}(\mu(x)-\mu_y) \\
&\quad+\frac{\nu}{\nu+n-2}\,\mathrm{Tr}\left(\Sigma_y^{-1}\Sigma_\phi(x)\right)-\frac{\nu C_1}{C_2}|\Sigma_\phi(x)|^{-\frac{\gamma}{2(1+\gamma)}}\Bigg],
\end{aligned}$$

and by taking $\Sigma_x=\sigma^2 I$ and $\Sigma_y=I$, we obtain :

$$\mathcal{L}_(\gamma,y)=\mathbb{E}_x\left[\frac{\mathbb{E}_z\left[||x-\mu_\theta(z)||^2\right]}{\sigma^2}+||\mu(x)-\mu_y)||^2+\frac{\nu\,\mathrm{Tr}\left(\Sigma_\phi(x)\right)}{\nu+n-2}-\frac{\nu C_1}{C_2}|\Sigma_\phi(x)|^{-\frac{\gamma}{2(1+\gamma)}}\right].$$

C Sampling distribution variance derivation

In this section, we present the derivation of τ^2 used in the sampling of t^3VAE and C-t^3VAE model. We present only the derivation for the C-t^3-VAE and it is identical to the one for the t^3-VAE since the former model is a generalization of the later.

First, we simplify the divergence $\mathcal{D}(q\|p^\star)$:

$$\mathcal{D}(q\|p^\star) = -\frac{C_{\nu+n,m}^{\frac{\gamma}{1+\gamma}}}{\gamma}\left(1+\frac{m}{\nu+n-2}\right)^{-\frac{\gamma}{1+\gamma}}\left[-\left(1+\nu^{-1}n\right)^{\frac{\gamma m}{2(1+\gamma)}}|\Sigma_\phi(x)|^{-\frac{\gamma}{2(1+\gamma)}}\left(1+\frac{m}{\nu+n-2}\right)\right.$$
$$\left.+|\tau^2 I|^{-\frac{\gamma}{2(1+\gamma)}}\times\left(1+\frac{\operatorname{Tr}\left(\tau^{-2}\left(1+\nu^{-1}n\right)^{-1}\Sigma_\phi(x)\right)}{\nu+n-2}+\frac{\tau^{-2}}{\nu+n}\|\mu(x)-\mu_y)\|^2\right)\right]$$

Here, we use the fact that $|\alpha A|^\delta = \alpha^{\delta n}|A|^\delta$ where n is the dimension of the square A matrix. Also, we use $\operatorname{Tr}(\alpha A) = \alpha\operatorname{Tr}(A)$. After simplification and rearranging we get :

$$\begin{aligned}\mathcal{D}(q\|p^\star) &= -\frac{1}{\gamma}C_{\nu+n,m}^{\frac{\gamma}{1+\gamma}}\left(1+\frac{m}{\nu+n-2}\right)^{-\frac{\gamma}{1+\gamma}}\left[-\left(1+\nu^{-1}n\right)^{\frac{\gamma m}{2(1+\gamma)}}|\Sigma_\phi(x)|^{-\frac{\gamma}{2(1+\gamma)}}\left(1+\frac{m}{\nu+n-2}\right)\right.\\
&\left.+\tau^{\frac{-\gamma m}{1+\gamma}}\left(1+\frac{\tau^{-2}\left(1+\nu^{-1}n\right)^{-1}}{\nu+n-2}\operatorname{Tr}\left(\Sigma_\phi(x)\right)+\frac{\tau^{-2}}{\nu+n}\|\mu(x)-\mu_y)\|^2\right)\right]\\
&= -\frac{1}{\gamma}C_{\nu+n,m}^{\frac{\gamma}{1+\gamma}}\left(1+\frac{m}{\nu+n-2}\right)^{-\frac{\gamma}{1+\gamma}}\left[-\left(1+\nu^{-1}n\right)^{\frac{\gamma m}{2(1+\gamma)}}|\Sigma_\phi(x)|^{-\frac{\gamma}{2(1+\gamma)}}\left(1+\frac{m}{\nu+n-2}\right)\right.\\
&\left.+\frac{1}{\nu+n}\tau^{-2-\frac{\gamma m}{1+\gamma}}\left(\kappa+\frac{\nu}{\nu+n-2}\operatorname{Tr}\left(\Sigma_\phi(x)\right)+\|\mu(x)-\mu_y)\|^2\right)\right]\\
&= -\frac{1}{\gamma}C_{\nu+n,m}^{\frac{\gamma}{1+\gamma}}\left(1+\frac{m}{\nu+n-2}\right)^{-\frac{\gamma}{1+\gamma}}\frac{1}{\nu+n}\tau^{-2-\frac{\gamma m}{1+\gamma}}\left[-(\nu+n)\tau^{2+\frac{\gamma m}{1+\gamma}}\left(1+\nu^{-1}n\right)^{\frac{\gamma m}{2(1+\gamma)}}\right.\\
&\left.|\Sigma_\phi(x)|^{-\frac{\gamma}{2(1+\gamma)}}\left(1+\frac{m}{\nu+n-2}\right)+\kappa+\frac{\nu}{\nu+n-2}\operatorname{Tr}\left(\Sigma_\phi(x)\right)+\|\mu(x)-\mu_y)\|^2\right],\end{aligned}$$

with:

$$\kappa = \tau^2(\nu+n).$$

Then, we match the result to the loss function in Eq. (9) to get :

$$\tau^{2+\frac{\gamma m}{1+\gamma}}\left(1+\nu^{-1}n\right)^{\frac{\gamma m}{2(1+\gamma)}+1}\left(1+\frac{m}{\nu+n-2}\right) = \frac{C_1}{C_2}.$$

Moreover, we have :

$$\begin{aligned}\frac{C_1}{C_2} &= C_{\nu+n,m}^{\frac{\gamma}{1+\gamma}}\left(1+\frac{n}{\nu}\right)^{\frac{\gamma m}{2(1+\gamma)}}\left(1+\frac{m}{\nu+n-2}\right)^{\frac{1}{1+\gamma}}C_{\nu,m+n}^{-\frac{\gamma}{1+\gamma}}\sigma^{\frac{n\gamma}{1+\gamma}}\left(1+\frac{m+n}{\nu-2}\right)^{\frac{\gamma}{1+\gamma}}\\
&= \sigma^{\frac{n\gamma}{1+\gamma}}C_{\nu+n,m}^{\frac{\gamma}{1+\gamma}}C_{\nu,m+n}^{-\frac{\gamma}{1+\gamma}}\left(1+\frac{n}{\nu}\right)^{\frac{\gamma m}{2(1+\gamma)}}\left(1+\frac{m}{\nu+n-2}\right)^{\frac{1}{1+\gamma}}\left(1+\frac{m+n}{\nu-2}\right)^{\frac{\gamma}{1+\gamma}}\\
&= \sigma^{\frac{n\gamma}{1+\gamma}}C_{\nu,n}^{\frac{-\gamma}{1+\gamma}}\left(1+\frac{m}{\nu+n-2}\right)^{\frac{1}{1+\gamma}}\left(1+\frac{m+n}{\nu-2}\right)^{\frac{\gamma}{1+\gamma}}.\end{aligned}$$

Consequently we obtain :

$$\tau^{2+\frac{\gamma m}{1+\gamma}}\left(1+\nu^{-1}n\right)^{\frac{\gamma m}{2(1+\gamma)}+1}\left(1+\frac{m}{\nu+n-2}\right) = \sigma^{\frac{n\gamma}{1+\gamma}}C_{\nu,n}^{\frac{-\gamma}{1+\gamma}}\left(1+\frac{m}{\nu+n-2}\right)^{\frac{1}{1+\gamma}}\left(1+\frac{m+n}{\nu-2}\right)^{\frac{\gamma}{1+\gamma}}$$

$$\begin{aligned}
\tau^{2+\frac{\gamma m}{1+\gamma}} &= \sigma^{\frac{n\gamma}{1+\gamma}} C_{\nu,n}^{\frac{-\gamma}{1+\gamma}} \left(1+\nu^{-1}n\right)^{-\frac{\gamma m}{2(1+\gamma)}-1} \left(1+\frac{m}{\nu+n-2}\right)^{-\frac{\gamma}{1+\gamma}} \left(1+\frac{m+n}{\nu-2}\right)^{\frac{\gamma}{1+\gamma}} \\
&= \sigma^{\frac{n\gamma}{1+\gamma}} C_{\nu,n}^{\frac{-\gamma}{1+\gamma}} \left(1+\nu^{-1}n\right)^{-\frac{\gamma m}{2(1+\gamma)}-1} \left(\frac{\nu+n-2}{\nu-2}\right)^{\frac{\gamma}{1+\gamma}} \\
&= \left(1+\nu^{-1}n\right)^{-\frac{\gamma m}{2(1+\gamma)}-1} \left(\sigma^n C_{\nu,n}^{-1} \quad \frac{\nu+n-2}{\nu-2}\right)^{\frac{\gamma}{1+\gamma}}.
\end{aligned}$$

Hence, we get :

$$\tau^2 = \left(1+\nu^{-1}n\right)^{-1} \left(\sigma^n C_{\nu,n}^{-1} \quad \frac{\nu+n-2}{\nu-2}\right)^{-\frac{2}{\nu+n-2}}.$$

which is the form of τ^2 we report in Eq. (8).

D Experimental setup

D.1 Datasets

We conduct experiments on three datasets notably SVHN-LT (Netzer et al., 2011), CIFAR100-LT (Krizhevsky, 2009; Cao et al., 2019) and CelebA (Liu et al., 2015) each chosen to highlight different challenges related to generative modeling under class imbalance and varying visual complexity.

- **SVHN-LT :** The Street View House Numbers (SVHN) dataset (Netzer et al., 2011) is composed of real-world digit images collected from Google Street View. It contains more than 600,000 labeled digits (0–9) with size 32×32, complex backgrounds, and diverse illumination conditions. The digits in SVHN are not centered or uniformly scaled, which makes the dataset considerably more challenging. Figure 6 shows sample images of this dataset.

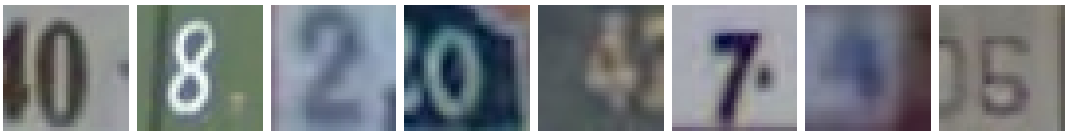


Figure 6: Sample images from the SVHN dataset (Netzer et al., 2011).

  However, as this dataset is naturally imbalanced, we balance the number of images across classes to have full control over the imposed imbalance ratio. In Table 2 we provide the number of images present in the dataset before balancing.

Class	0	1	2	3	4	5	6	7	8	9
Train set	4948	13861	10585	8497	7458	6882	5727	5595	5045	**4656**
Test set	1744	5099	4149	2882	2523	2384	1977	2019	1660	**1595**

Table 2: The Number of images in the SVHN dataset for the train and test sets before balancing. The value in bold is the one used to balance the dataset.

  For both training and testing, we crop each class to the minimum number of samples available across all classes. The only data augmentation applied is a random horizontal flip with 50% probability.

- **CIFAR100-LT :** The CIFAR100 dataset (Krizhevsky, 2009) consists of 60,000 color images of size 32×32 pixels, evenly distributed across 100 object categories. Each category contains 600 images, split into 500 training samples and 100 test samples. Figure 7 shows sample images after preprocessing.

Figure 7: Sample images from the CIFAR100 dataset (Krizhevsky, 2009).

We use the dataset in its entirety without class filtering. As with SVHN-LT, we apply a random horizontal flip with 50% probability for data augmentation.

- **CelebA :** The CelebFaces Attributes Dataset (CelebA) (Liu et al., 2015) contains 202,599 color images of celebrity faces at a resolution of 178×218. Each image is annotated with 40 binary facial attributes, such as Mustache, Smiling, and Young. This dataset exhibits significant variability in pose, expression, and illumination while providing high resolution images. We preprocess the images of this dataset by cropping the central region to 160×160 and resizing to 128×128 using bilinear interpolation. Figure 8 illustrates sample images after preprocessing. Also, we adhere to the original training, validation, and test splits provided by the dataset.

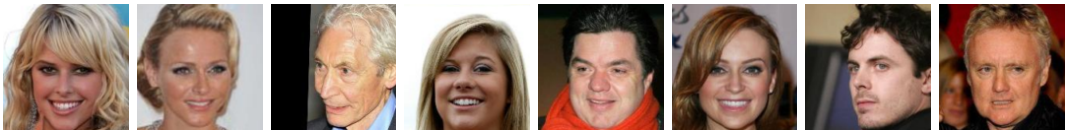


Figure 8: Sample images from the CelebA dataset after preprocessing (Liu et al., 2015).

	Mustache	No Mustache	Young	Old	Male	Female	Smiling	Not Smiling
Train set	6642	156128	126788	35982	68261	94509	78080	84690
Test set	722	19190	15114	4848	7715	12247	9987	9975

Table 3: Number of images in the CelebA dataset for the train and test sets for the Mustache, Young, Male and Smiling attributes.

Given CelebA's multi-attribute structure, we select four binary attributes—Mustache, Young, Male, and Smiling—for training and evaluation, treating each attribute and its negation as distinct classes. These attributes were chosen due to their varying imbalance ratios, defined as the frequency of the majority class divided by the minority. The resulting ratios are 25 for Mustache, 3.5 for Young, 1.4 for Male, and 1 for Smiling. This enables the study of imbalance effects in generative modeling. We also balance the test sets by downsampling the larger class for each attribute. In Table 3 we report the number of images per selected attribute in the dataset.

D.2 Model Architecture

Our encoder-decoder models follow a modular block design. Each encoder block consists of a convolutional layer, followed by 2D batch normalization and ReLU activation. Decoder blocks mirror this structure but replace convolutional layers with transposed convolutions.

- **SVHN-LT and CIFAR100-LT :** Encoders consist of four convolutional blocks with channels {64, 128, 256, 512}, followed by two linear layers for estimating mean and covariance. The decoder uses three transposed convolutional blocks with channel sizes {128, 64, 32}, ending with a three-channel convolution and Sigmoid activation.
- **CelebA :** The CelebA encoder includes six convolutional blocks with channels {64, 128, 256, 512, 512, 512}, ending with two linear layers. The decoder has six transposed convolutional layers with channels {512, 512, 256, 128, 64, 32}, followed by a final convolutional layer and Sigmoid activation.

D.3 Training Details

All models are trained using the AdamW optimizer with a learning rate of 10^{-3} for 150 epochs. We use a batch size of 64 for SVHN-LT and CIFAR100-LT, and 128 for CelebA.

E Hyperparameter Tuning

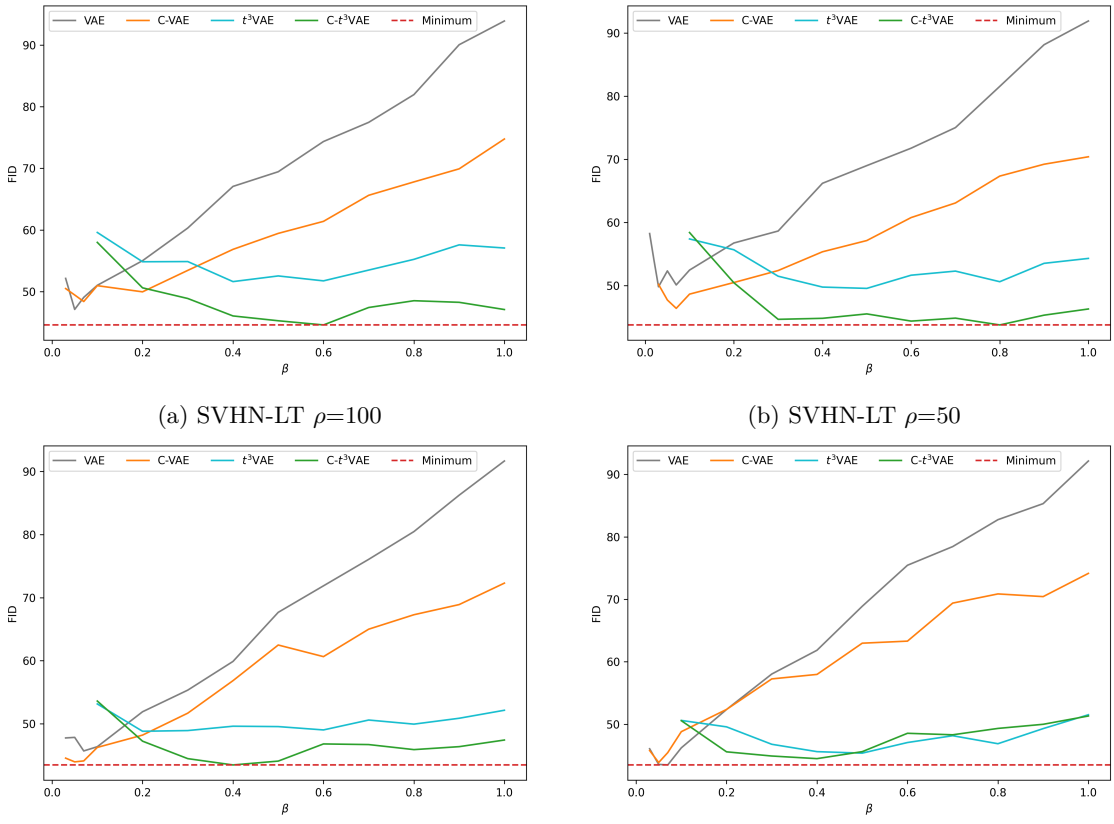
We present the hyperparameter tuning process used across all evaluated models. We first optimize β, then ν, and finally τ, yielding the models' results reported in Table 1.

E.1 β Optimization

We perform a hyperparameter study over β for all tested models. Unless otherwise noted, we use the theoretically derived τ^2 and set $\nu = 10$.

E.1.1 On the SVHN-LT dataset

As shown in Figure 9, the optimal β values for Student's t models lies in the range $\beta \in [0.4, 0.6]$ whereas it lies in the $\beta \in [0.05, 0.07]$ range for Gaussian-based models. This is because the regularization term in the γ-power divergence loss is ten times larger than the KL divergence. Figure 9 also shows that FID performance is highly sensitive to β in the Gaussian setting, requiring careful tuning which is not the case for Student's t based models. Finally, C-t^3VAE achieves the best FID surpassing the t^3VAE and the C-VAE for all imbalance settings.



(a) SVHN-LT ρ=100

(b) SVHN-LT ρ=50

(c) SVHN-LT ρ=10

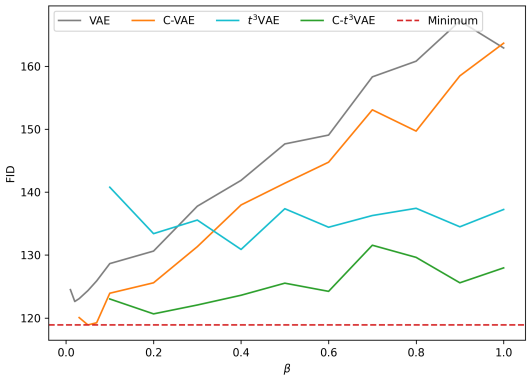
(d) SVHN-LT ρ=1

Figure 9: Variability of the FID as a function of the β hyperparameter for the VAE, C-VAE, t^3VAE and C-t^3VAE on the SVHN-LT dataset.

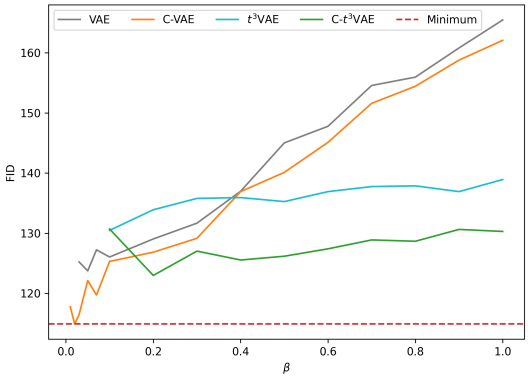
E.1.2 On the CIFAR100-LT dataset

From Figure 10, we observe that Student's t models obtain the best performance in terms of FID at $\beta = 0.2$ for CIFAR100-LT dataset. However, for the Gaussian-based models, the optimal value is much lower with $\beta \in [0.02, 0.05]$. The reason for this is that on this dataset too the KL regularization term is ten times smaller than the regularization terms present in the γ-power divergence loss. Additionally, we notice that C-VAE performs slightly better, likely due to the complexity of the dataset preventing full convergence to the imposed latent distribution. We further investigate this hypothesis in the τ analysis below.

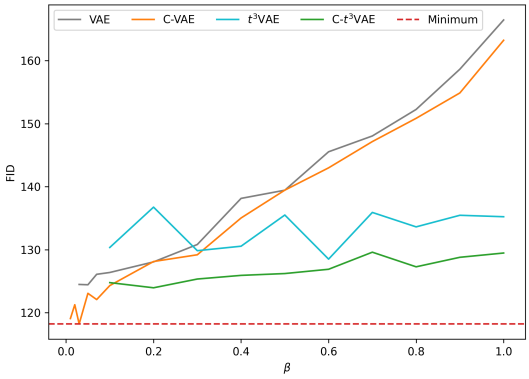
(a) CIFAR100-LT ρ=100



(b) CIFAR100-LT ρ=50



(c) CIFAR100-LT ρ=10

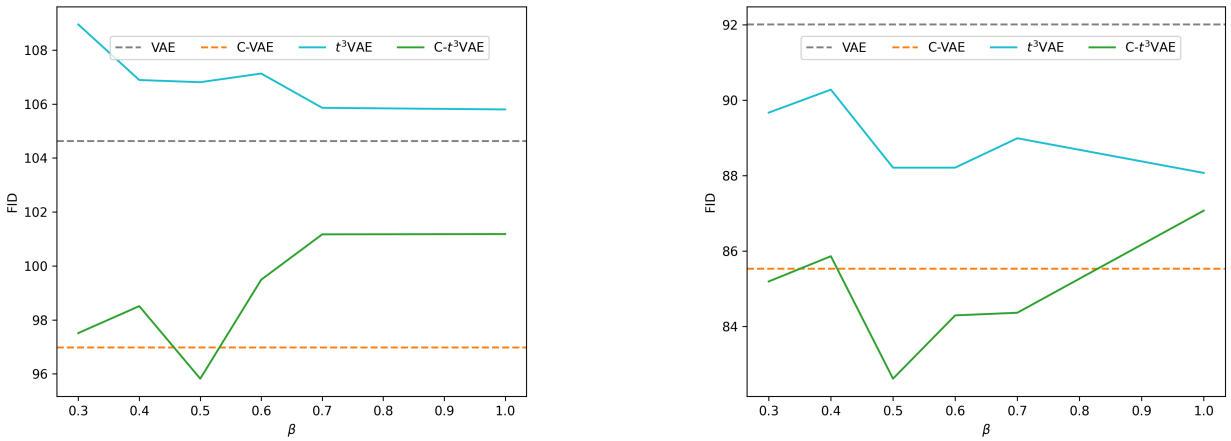


(d) CIFAR100-LT ρ=1

Figure 10: Variability of the FID as a function of the β hyperparameter for the VAE, C-VAE, t^3VAE and C-t^3VAE on the CIFAR100-LT dataset.

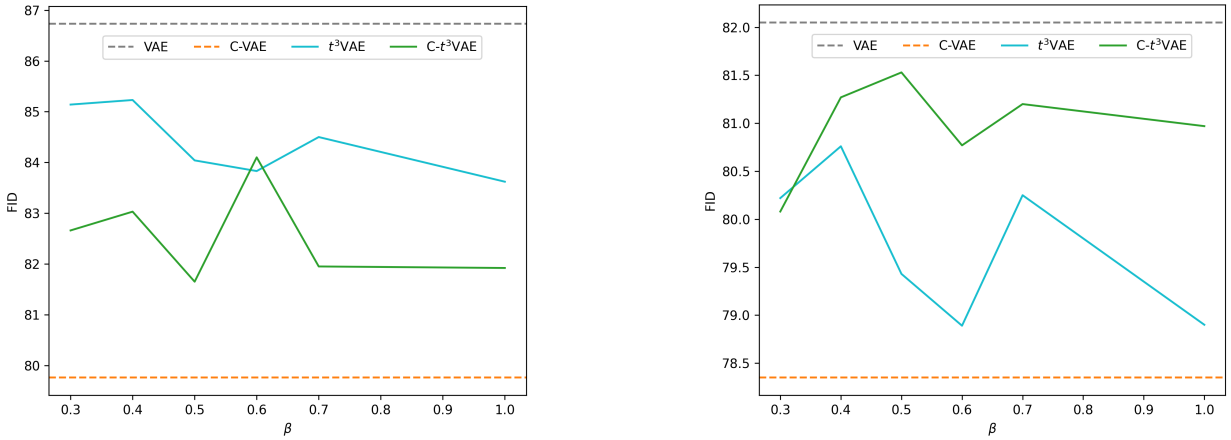
E.1.3 On the CelebA dataset

For CelebA, we optimize β exclusively for Student's t models, setting $\beta = 0.1$ for Gaussian variants. Table 1 indicates that β has minimal impact on CelebA's FID, unlike the trends seen in SVHN-LT and CIFAR100-LT. Therefore, extensive tuning for Gaussian models is omitted. Figure 11 suggests that Student's t models share this robustness to β variations, likely driven by the dataset's limited intra-class variability.



(a) CelebA - Mustache

(b) CelebA - Young



(c) CelebA - Male

(d) CelebA - Smiling

Figure 11: Variability of the FID as a function of the β hyperparameter for the t^3VAE and C-t^3VAE on the CelebA dataset. The horizontal lines for the VAE and C-VAE models are for the best performing model between $\beta = 0.1$ and $\beta = 1$.

E.2 ν Optimization

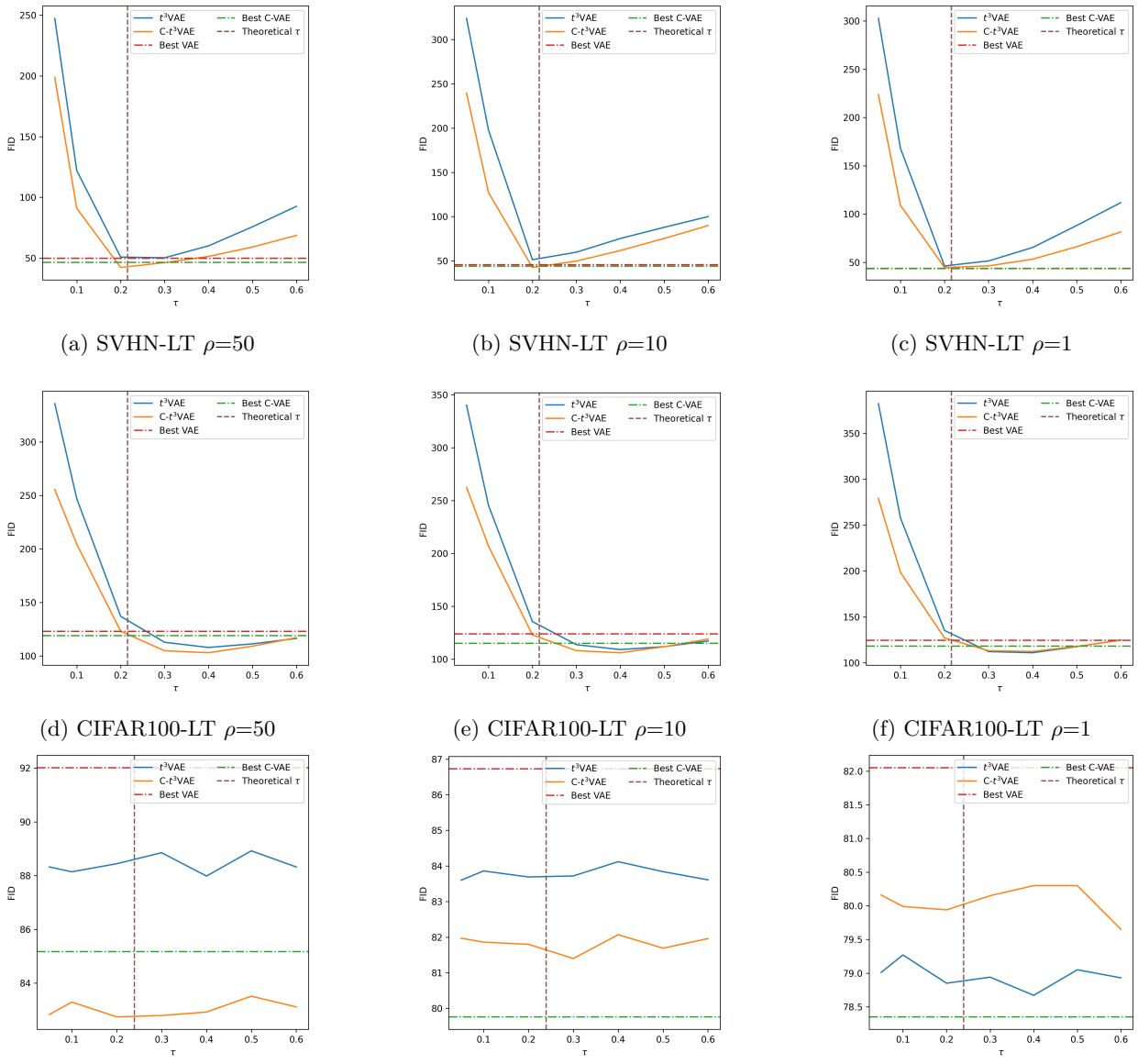
Table 4 presents a sensitivity analysis of the degrees of freedom parameter ν for the C-t^3VAE on SVHN-LT and CIFAR100-LT, using the optimal β from the previous study. Consistent with prior work (Kim et al., 2024), we find that $\nu = 10$ yields robust performance on average, though slight gains can be achieved by fine-tuning within the range $[2.5, 20]$. Ultimately, however, the generative FID remains relatively insensitive to variations in ν, corroborating the findings of (Kim et al., 2024) regarding reconstruction quality.

	SVHN-LT				CIFAR100-LT			
ν	100	50	10	1	100	50	10	1
2.1	45.50	44.51	42.96	46.23	121.28	122.03	121.93	123.41
2.5	45.76	43.96	45.81	45.40	119.15	120.19	**120.10**	124.83
5	44.89	**42.60**	45.03	46.33	120.52	123.21	124.29	123.71
10	44.59	44.37	43.48	**44.49**	**119.83**	120.65	122.96	**123.95**
20	**44.02**	43.89	**42.01**	44.75	121.48	**118.41**	124.58	126.13
50	48.03	46.39	43.59	45.57	**119.58**	126.36	124.38	127.48
100	45.97	44.63	43.74	47.52	123.26	122.90	127.42	125.67

Table 4: Variability of the FID as a function of the standard deviation ν for the C-t^3VAE model.

E.3 τ Optimization

In this section, we evaluate the effect of the τ parameter on the SVHN-LT, CIFAR100-LT and CelebA datasets for all imbalance ratios while setting β and ν to their previously optimized values. As shown in Figure 12, the optimal τ for SVHN-LT aligns closely with our theoretical prediction. In contrast, CIFAR100-LT consistently benefits from a larger $\tau = 0.4$, yielding improved FID across all imbalance settings and outperforming C-VAE. On CelebA, τ has minimal impact and the most likely value is $\tau \approx 0.3$. The analytically derived τ provides a principled initialization grounded in divergence geometry, while dataset-specific deviations reflect encoder–decoder capacity limits rather than theoretical inconsistency, highlighting the interaction between prior heaviness and representation complexity.



(a) SVHN-LT ρ=50 (b) SVHN-LT ρ=10 (c) SVHN-LT ρ=1

(d) CIFAR100-LT ρ=50 (e) CIFAR100-LT ρ=10 (f) CIFAR100-LT ρ=1

(g) CelebA Young (h) CelebA Male (i) CelebA Smiling

Figure 12: Variability of the FID as a function of the standard deviation τ^2 for the t^3VAE and C-t^3VAE. In horizontal dashed lines is the FID value of the best performing VAE and C-VAE on each dataset. In vertical dashed lines is the theoretically identified value of τ.

F Per-Class Evaluation

In this section, we assess the conditional models' per-class Recall, Precision, and F1 metrics under all imbalance settings and for all tested datasets after optimization of all hyper-parameters.

From the following figures in Table 5 and 6, we see that the C-t^3VAE consistently improves Recall and mode coverage in highly imbalanced settings with $\rho = 100$ and $\rho = 50$. This comes at a minor Precision cost but results in significantly better F1 scores across most classes. However, on balanced or mildly imbalanced datasets, its performance remains competitive with Gaussian-based models. This observation is valid for both the SVHN-LT and CIFAR100-LT but is more pronounced on the later.

Table 5: Per-class generative metrics on SVHN-LT after optimization of β, ν and τ hyper-parameters.

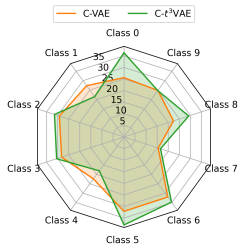
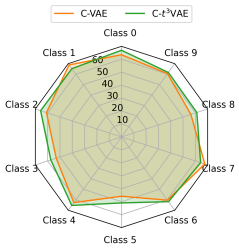
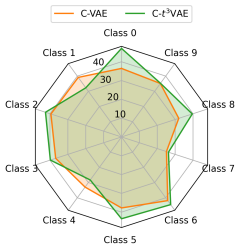
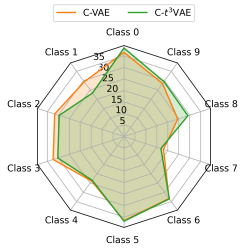
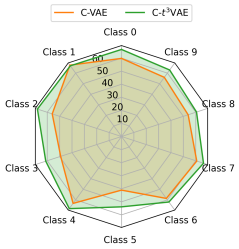
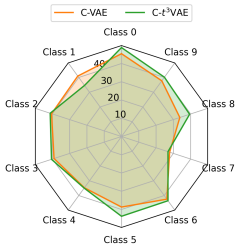
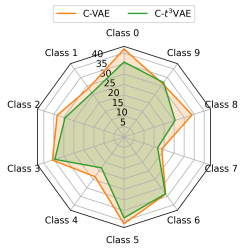
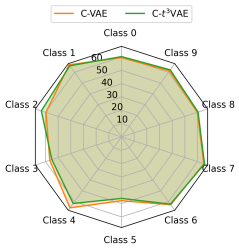
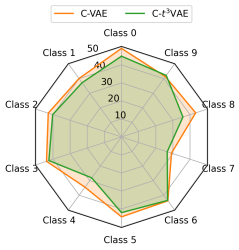
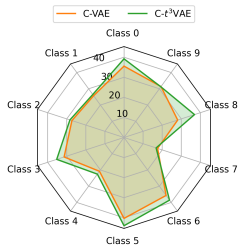
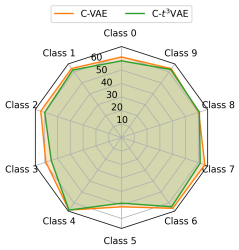
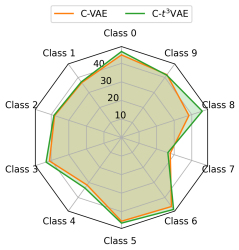
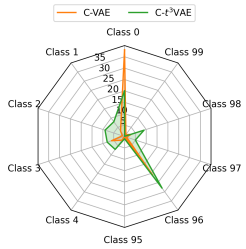
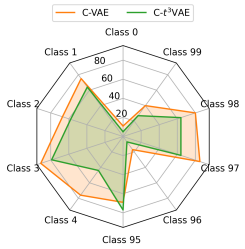
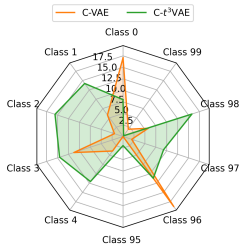
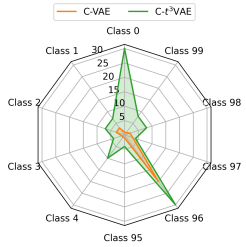
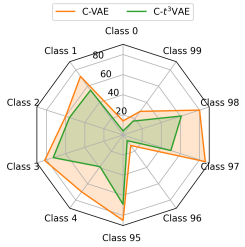
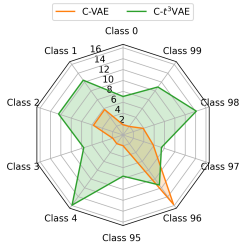
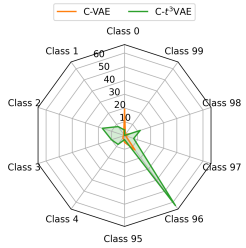
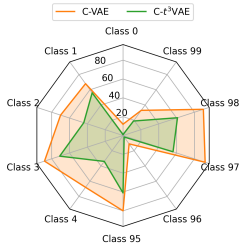
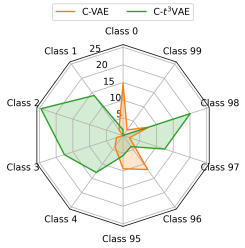
ρ	Recall	Precision	F1 score
100			
50			
10			
1			

Table 6: Per-class generative metrics on CIFAR100-LT after optimization of β, ν and τ hyper-parameters, we focus on the top 5 head and tail classes.

ρ	Recall	Precision	F1 score
100			
50			
10			
1