# $\lambda$-VAE: Variance Equalization for Posterior Collapse

**Anonymous authors**
**Paper under double-blind review**

## Abstract

Variational Autoencoders (VAEs) frequently suffer from posterior collapse, a failure mode in which the approximate posterior converges to the prior, rendering the latent code uninformative. Despite extensive research, a unified account of why collapse occurs has remained an open question. We identify and formalize two logically independent but coupled causes. *Gradient imbalance* occurs when the decoder's reconstruction signal vanishes faster than the $\mathbb{KL}$ regularization pressure as the posterior widens. *Information gap* occurs when the stochastic sampling step discards a substantial fraction of the encoder's computed representation, attenuating decoder sensitivity and making collapse inexpensive. Both causes share the same collapse trajectory, and we show that the information gap is algebraically equivalent to marginal mismatch between the aggregate posterior and the prior, unifying two pathologies previously treated as separate.We introduce $\lambda$-VAE, which resolves both causes through a single modification to the reparameterization step: the sampling noise is scaled by per-dimension exponent, while the $\mathbb{KL}$ penalty retains the original posterior variance. This asymmetry shifts the stable training attractor away from the degenerate collapsed state, driving all latent dimensions toward the same equilibrium – a mechanism we term *variance equalization*. A closed-form optimal exponent per dimension follows from a net information gain objective, with a single hyperparameter controlling the reconstruction–generation tradeoff. We validate on standard benchmarks (Binary MNIST, Binary Omniglot, CIFAR-10, CelebA-64), showing consistent reductions in collapsed dimensions, information capacity gains of up to 2.8×, and reconstruction quality improvements of up to +0.33 BPD.

## 1 Introduction

Variational Autoencoders (VAEs) Kingma & Welling (2013); Rezende et al. (2014) have become a foundational tool for learning latent representations, with applications in image generation Razavi et al. (2019b), natural language processing Bowman et al. (2015), multimodal modelling Shi et al. (2019), and semi-supervised learning Kingma et al. (2014). VAE is one of many variational inference frameworks that use approximate posterior for maximizing data likelihood Neal & Hinton (1998); McLachlan & Krishnan (2007); Minka (2013); Jordan et al. (1999). It is, however, one of the first to use *amortization* to estimate approximate posterior; an encoder learns to map each input to a posterior distribution over a latent space, and a decoder reconstructs the input from a sample. This is a highly efficient and scalable approach, but it is prone to a well-documented failure mode called *posterior collapse* Bowman et al. (2015); Chen et al. (2016); Lucas et al. (2019), in which the approximate posterior converges to the prior partially or fully. When collapse occurs, the decoder learns to reconstruct data from the prior. In its extreme form, the mutual information between data and latent variables goes to zero ($I(X; Z) \approx 0$), and the latent space loses all its representational value. This failure is most severe with high-capacity decoders Bowman et al. (2015); Razavi et al. (2019b), such as those with autoregressive or deep convolutional architectures, but it occurs to varying degrees in most practical VAE training runs.

Extensive research has proposed different solutions for posterior collapse; re-weighting the Kullback-Leibler ($\mathbb{KL}$) divergence term Higgins et al. (2016); Fu et al. (2019), imposing minimum $\mathbb{KL}$ floors Chen et al. (2016); Razavi et al. (2019a), matching the aggregate posterior explicitly Zhao et al. (2019); Tolstikhin et al. (2017), enriching the posterior family Rezende & Mohamed (2015); Kingma et al. (2016), and adjusting training

dynamics He et al. (2019). These methods target different manifestations of the problem and share a common limitation: they act on the training objective and propose a global solution, while lacking a unified account on the causes of collapse.

In this paper, we provide a formal account of why posterior collapse occurs and build a targeted solution on top of it. Our starting observation is simple: in a standard VAE, the $\mathbb{KL}$ term selectively pushes low-signal dimensions toward $\sigma_i = 1$ while leaving high-signal dimensions active, producing a *polarized* distribution of posterior variances. This polarization is both the signature of emerging collapse and a driver of further collapse. Hence, the right intervention is not to globally modify the training objective but to counteract the polarization directly, which we formalize and validate in subsequent sections.

**Summary of Contributions:**

1. We prove that posterior collapse has two logically independent but coupled causes: *gradient imbalance* (Proposition 1), in which the reconstruction gradient vanishes before the $\mathbb{KL}$ restoring force as the posterior widens, and *information gap* (Proposition 2), in which the stochastic bottleneck discards encoder signal, making collapse inexpensive. We further show that the information gap is algebraically equivalent to marginal mismatch between the aggregate posterior and the prior (Eq. 8).

2. Replacing the reparameterization noise $\sigma\epsilon$ with $\sigma^\lambda\epsilon$ while retaining the $\mathbb{KL}$ penalty on the original $\sigma^2$ resolves both causes simultaneously. The asymmetry drives all latent dimensions toward the same equilibrium – *variance equalization* – without modifying the training objective or adding parameters. A closed-form optimal per-dimension exponent follows from a net information gain objective (Proposition 3), with a single hyperparameter $\delta$ controlling the reconstruction–generation tradeoff.

3. On binarized benchmarks, $\lambda$-VAE reduces collapsed dimensions from 16 to 1 on Binary MNIST and 13 to 0 on Binary Omniglot while improving reconstruction quality. On RGB images, SNR information capacity grows by up to 2.8× and ELBO-BPD improves by +0.33 on CIFAR-10. A PixelCNN experiment reveals a fundamental limitation of BPD as a collapse diagnostic: two models with near-identical BPD (3.518 vs. 3.494) differ by 6.2× in the decoder capacity allocated to the latent code, a disparity invisible to BPD alone.

## 2 Variational Autoencoders

VAE Kingma & Welling (2013); Rezende et al. (2014) is one of the earlier approaches to use *amortization* and stochastic gradient estimation in large-scale datasets via the Evidence Lower Bound (ELBO), written as

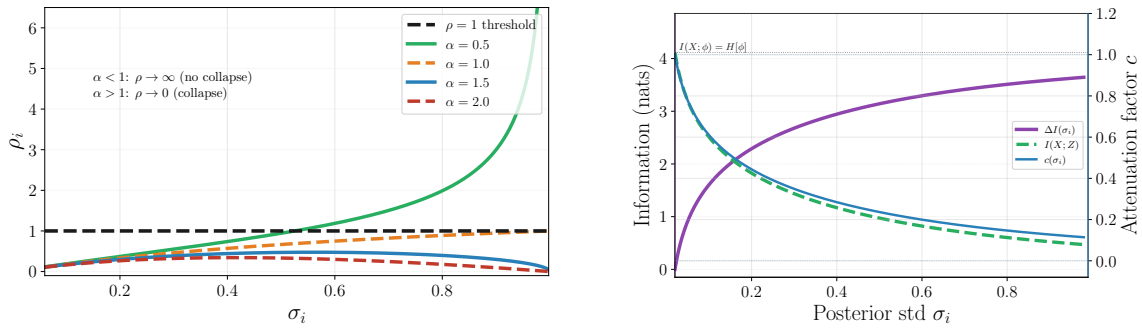
$$\mathcal{L} = \mathbb{E}_{q_\psi(z|x)}[\log p_\theta(x|z)] - \mathbb{KL}(q_\psi(z|x) \,\|\, p(z)) \,, \tag{1}$$

with $q_\psi(z|x) = \mathcal{N}(\mu(x), \mathrm{diag}(\sigma^2(x)))$, $p(z) = \mathcal{N}(0, I)$, and sampling via $z = \mu(x) + \sigma(x)\epsilon$, $\epsilon \sim \mathcal{N}(0, I)$. The Gaussian assumption on the posterior and prior makes the $\mathbb{KL}$ term computable in a closed form; $\mathbb{KL}_i = \frac{1}{2}(\mu_i^2 + \sigma_i^2 - \log\sigma_i^2 - 1)$, henceforth we work per latent dimension $i$ and suppress the subscript where unambiguous. In Hoffman & Johnson (2016), the $\mathbb{KL}$ term in ELBO is decomposed and the overall loss is written as

$$\mathcal{L} = \mathbb{E}[\log p_\theta(x|z)] - I_q(X;Z) - \mathbb{KL}(q_\psi(z)\|p(z)), \tag{2}$$

where $I_q(X;Z)$ is the mutual information, and $\mathbb{KL}(q_\psi(z)\|p(z))$ is the marginal mismatch. Hence, minimizing the $\mathbb{KL}$ term in Eq. 1 reduces both $I_q(X;Z)$ and the marginal; the optimizer cannot distinguish between them, which is why collapse is an unintended consequence. That is, at $\mathbb{KL} \approx 0$ in Eq. 1, the latent code $z$ becomes statistically independent of the input $x$, rendering the encoder information useless in the signal reconstruction.

In general, ELBO creates an information bottleneck that encourages encoding the least amount of information subject to reconstruction error Tishby et al. (2000). Posterior collapse, however, is an extreme case in which the model achieves sufficient reconstruction accuracy while ignoring all input signal, $I_q(X;Z) \approx 0$ – often




(a) Gradient imbalance ($\rho_i$ vs $\sigma_i$)

(b) Information gap ($\Delta I$, $I(X;Z)$, $c$ vs $\sigma_i$)

Figure 1: **Posterior collapse causes.** *(a)* Gradient ratio $\rho_i$ under the decay model $g_i^\sigma = g_0(1-\sigma_i)^\alpha$, $g_0 = 2$. For $\alpha > 1$ (Proposition 1), $\rho_i \to 0$ as $\sigma_i \to 1$: the reconstruction signal vanishes faster than the $\mathbb{KL}$ restoring force, making the collapsed state the only stable fixed point. For $\alpha < 1$, $\rho_i$ is no longer a reliable indicator: it stays above 1 even when the dimension has collapsed, because the reconstruction gradient from the decoder decays much more slowly in this regime. *(b)* Mutual information $I(X;Z)$, information gap $\Delta I$, and attenuation factor $c = I(X;Z)/I(X;\phi)$ using the Gaussian channel model with $S_i = 1.5$. As $\sigma_i$ increases towards 1: $I(X;Z)$ falls, $\Delta I$ grows toward $I(X;\phi)$, and $c \to 0$. Hence, collapse becomes much cheaper (Proposition 2).

observed and measured per latent dimension. In Section 3, we will expand on both data and model specific conditions that enable VAE training to find solutions that fail to encode details about the input while achieving high reconstruction accuracy.

## 3 Posterior Collapse: Two Causes

Posterior collapse in trained models manifests as near-zero per-dimension $\mathbb{KL}$, a low count of active latent dimensions, and $I_q(X;Z) \approx 0$. In the subsequent sections, we present two principal causes for posterior collapse and show thier relationship to $\mathbb{KL}$ and mutual information.

### 3.1 Gradient Imbalance

Differentiating Eq. 1 with respect to $\sigma_i$ via the reparameterisation $z_i = \mu_i + \sigma_i\epsilon_i$, we have

$$\frac{\partial \mathcal{L}}{\partial \sigma_i} = \mathbb{E}_{p(x),\,\epsilon_i}\big[\nabla_{z_i} \log p_\theta(x|z) \cdot \epsilon_i\big] + \frac{1}{\sigma_i} - \sigma_i. \tag{3}$$

The $\mathbb{KL}$ restoring force (second term) is positive for $\sigma_i < 1$, zero at $\sigma_i = 1$, and negative for $\sigma_i > 1$. As such, it always drives $\sigma_i$ toward 1 while looking for a stable solution in Eq. 3. When the first term, the reconstruction gradient which we denote by $g_i^\sigma$, vanishes then $\sigma_i = 1$ is the sole equilibrium. Hence, the dynamics between $g_i^\sigma$ and the $\mathbb{KL}$ restoring force determine whether a dimension remains active or collapses during training.

We define the *gradient ratio* $\rho_i = |g_i^\sigma|/|1/\sigma_i - \sigma_i|$ as the primary indicator of collapse. That is, when $\rho_i \ll 1$, the $\mathbb{KL}$ term dominates and the dimension collapses. The following proposition formalizes this understanding.

**Proposition 1** (Gradient Imbalance)**.** *Suppose the asymptotic decay condition for the reconstruction gradient is $|g_i^\sigma| = O((1-\sigma_i)^\alpha)$ as $\sigma_i \to 1^-$ with $\alpha > 1$. Then dimension $i$ collapses to the prior if and only if $\liminf_{t\to\infty} \rho_i(t) < \rho_{\text{crit}}$ for some finite $\rho_{\text{crit}} > 0$.*[1]

[1]The idealized gradient-balance value is $\rho_{\text{crit}} = 1$ (any non-collapsed equilibrium of $\partial\mathcal{L}/\partial\sigma_i$ satisfies $\rho_i \geq 1$). In practice, joint encoder–decoder optimization shifts the effective threshold; on CIFAR-10, for instance, the empirical $\rho_{\text{crit}} \approx 0.006$ separates active from collapsed dimensions with 99.8% accuracy. The value 1 therefore bounds the theoretical regime; the operative threshold is model and dataset specific.

Under this condition, $\rho_i \to 0$ as $\sigma_i \to 1$, making $\sigma_i = 1$ the sole stable fixed point. The decay exponent $\alpha$ characterizes the decoder's sensitivity to the latent code near collapse. The condition $\alpha > 1$ identifies decoders for which $\rho_i$ is a necessary and sufficient collapse signal. Autoregressive decoders satisfy this with $\alpha \gg 1$, explaining their high collapse rate Bowman et al. (2015) and empirically confirmed by the PixelCNN experiment in Section 6.4. The collapsed fixed point $\sigma_i^* = 1$ is locally stable with linearization rate $-2$; by contrast, $\mu_i^* = 0$ has rate $-1$, which is why $\sigma_i \to 1$ is the more reliable leading indicator of collapse. Full proof and details are in Appendix A, while illustration of the dynamics is shown in Figure 1a.

### 3.2 Information Gap

The encoder in a VAE, which estimates the posterior parameters, is a deterministic learned map $\phi(x) = (\mu(x), \log \sigma(x))$, while the latent $Z$ is obtained by stochastic reparameterization. This defines a Markov chain $X \to \phi(X) \to Z$, for which the data processing inequality Cover (1999) gives:

$$H[X] \;\geq\; I(X;\phi) \;\geq\; I(X;Z) \;\geq\; 0, \tag{4}$$

where $H(\cdot)$ denotes entropy of a random variable. Since $I(X;\phi) = H[\phi]$ (encoder is deterministic), the *information gap* is the information capacity that the sampling step loses, captured as

$$\Delta I \;=\; I(X;\phi) - I(X;Z) \;\geq\; 0. \tag{5}$$

The information gap widens if the dataset has low signal power (low variance in $\mu_i(X)$), which reduces $I(X;Z)$ as $\sigma \to 1$. More importantly, a large gap implies an attenuation factor that dampens decoder sensitivity to encoder change, facilitating collapse. The following proposition formalizes this.

**Proposition 2** (Information Gap)**.** *Let $\mathcal{R} = f(I(X;Z))$ be reconstruction quality measure for monotonically increasing $f$. The sensitivity of reconstruction quality to encoder capacity is attenuated as*

$$\frac{\partial \mathcal{R}}{\partial I(X;\phi)} = \frac{\partial \mathcal{R}}{\partial I(X;Z)} \cdot \left(1 - \frac{\partial \Delta I}{\partial I(X;\phi)}\right). \tag{6}$$

The attenuation factor $c = 1 - \partial \Delta I / \partial I(X;\phi) \approx I(X;Z)/I(X;\phi)$ (under the approximation that the two move proportionally during training) measures how much of the encoder's signal the decoder can access. The Gaussian channel model makes this concrete by treating $z_i = \mu_i(X) + \sigma_i \epsilon_i$ as transmitting signal $\mu_i$ through noise $\sigma_i \epsilon_i$, giving

$$I(X;Z_i) \;\approx\; \tfrac{1}{2} \log\left(1 + \frac{S_i}{\sigma_i^2}\right), \tag{7}$$

where $S_i = \mathrm{Var}[\mu_i(X)]$ is the signal power. High SNR (signal-to-noise ratio) gives $c \approx 1$, making collapse costly; low SNR gives $c \to 0$ and collapse exerts zero pressure on the decoder – it is free (Figure 1b). More importantly, under marginal matching where $q(z) \approx \mathcal{N}(0, \mathbf{I})$ the signal power $S_i \approx 1 - \sigma_i^2$, this reduces to $I(X;Z_i) \approx -\log \sigma_i$ as a bound to the accessible infomration which goes to zero as $\sigma \to 1$; result is further discussed in Appendix A.

Overall, the information gap decomposes into a reducible and an irreducible part. The irreducible floor is set by the data's intrinsic dimensionality: dimensions where the data provides no signal which contribute an unavoidable gap regardless of encoder quality. For instance, in binarized MNIST and Omniglot, corner pixels carry no information; when encoded into a high-dimensional latent space, the corresponding dimensions will naturally collapse. The reducible part is $\mathbb{KL}$-induced; since $\mathbb{KL}$ forces $S_i \approx 1 - \sigma_i^2$, any rise in $\sigma_i$ under regularization pressure directly suppresses $S_i$, creating an information gap a higher-SNR encoder could avoid. Intuitively, a lower information gap implies a more distributed representation – dimensions carry, ideally, independent rather than redundant signal – whereas a sparse representation concentrates signal in few overloaded dimensions and leaves the rest high-noise.

A direct consequence follows from the ELBO decomposition Hoffman & Johnson (2016). Substituting $I(X;Z) = I(X;\phi) - \Delta I$ into Eq. 2:

$$\mathbb{KL}(q_\psi(z) \| p(z)) = \mathbb{KL}(q_\psi(z|x) \| p(z)) - I(X;\phi) + \Delta I. \tag{8}$$

This identity holds at every training snapshot. Any increase in $\Delta I$ that is not compensated by a decrease in $\mathbb{KL}(q(z|x)\|p(z)) - I(X;\phi)$ manifests directly as growing marginal mismatch. Encoder attenuation and aggregate posterior misalignment are therefore algebraically equivalent: two descriptions of the same quantity, one from the sampling bottleneck and one from the aggregate distribution. Hence, the information gap serves as a cheap, per-batch proxy for marginal mismatch.

### 3.3 Coupling Between the Two Causes

The two causes are logically independent; gradient imbalance depends on decoder architecture ($\alpha$, and decoder capacity); the information gap depends on data complexity ($S_i$) and posterior width ($\sigma_i$). Either can occur without the other. However, both are driven by $\sigma_i \to 1$: gradient imbalance makes this trajectory dynamically inevitable once $\rho_i < \rho_{\text{crit}}$ under the asymptotic condition $\alpha > 1$; the information gap makes reaching $\sigma_i = 1$ less costly before that point.

Once gradient imbalance pushes $\rho_i$ below critical point, $\sigma_i$ begins to rise. For data with low latent complexity $S_i$, the rising $\sigma_i$ opens a substantial information gap, lowers the attenuation factor $c$, and weakens the reconstruction gradient $g_i^\sigma$ which reduces $\rho_i$ further and accelerates the rise in $\sigma_i$. The feedback cycle is closed and positive; gradient imbalance provides the initial push, and the information gap amplifies it when the encoder has high capacity. On complex data, the information gap remains shallow and the attenuation is weak; collapse is then driven primarily by gradient imbalance alone. In both cases, however, the polarization of $\sigma$ is the signature of the trajectory toward collapse, and the key point of intervention. In Section 4, we present an approach that actively counteracts this polarization, addressing both causes simultaneously.

## 4 $\lambda$-VAE: Variance Equalization

Consider the modification of the latent sampling step in ELBO as

$$z = \mu(x) + \sigma(x)^\lambda \cdot \epsilon, \qquad \lambda \geq 1, \tag{9}$$

while computing the $\mathbb{KL}$ term using the original variance $\sigma(x)^2$. Subsequently, the objective will be modified as

$$\mathcal{L}_\lambda = \mathbb{E}_{q_\lambda(z|x)}[\log p_\theta(x|z)] - \mathbb{KL}(q_\psi(z|x) \,\|\, p(z)) \,, \tag{10}$$

where $q_\lambda(z|x) = \mathcal{N}(\mu(x), \text{diag}(\sigma(x)^{2\lambda}))$ is the sampling distribution but the $\mathbb{KL}$ penalty uses $\sigma(x)^2$ and not $\sigma(x)^{2\lambda}$; this is $\lambda$-VAE.

$\lambda$-VAE introduces an asymmetry between the sampling and penalty terms. The decoder receives noise $\sigma_i^\lambda \epsilon$ (reduced relative to $\sigma_i \epsilon$ for $\sigma_i < 1$, $\lambda > 1$) while the $\mathbb{KL}$ penalty evaluates the original $\sigma_i^2$. This decoupling allows more information to flow through the bottleneck without inflating the $\mathbb{KL}$ penalty.

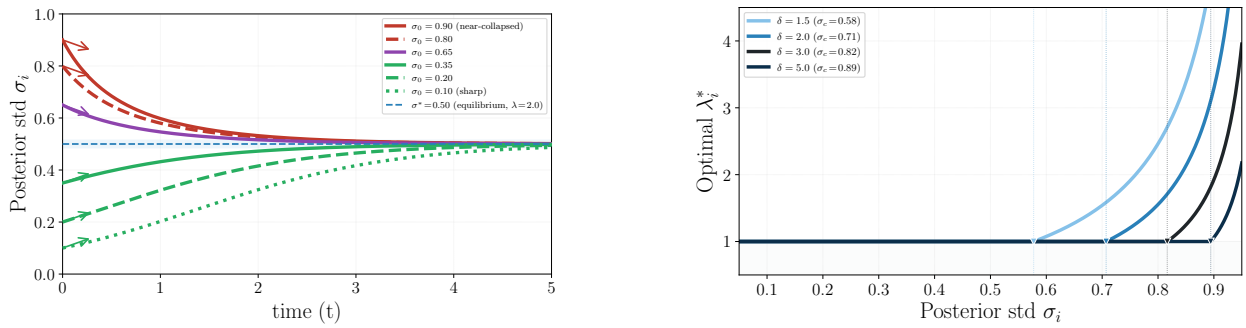
### 4.1 Gradient Rebalancing

The ELBO gradient with respect to $\sigma_i$ under $\lambda$-scaling is

$$\frac{\partial \mathcal{L}_\lambda}{\partial \sigma_i} = \lambda\, \sigma_i^{\lambda-1} \cdot g_i^\sigma + \frac{1}{\sigma_i} - \sigma_i. \tag{11}$$

Hence, the reconstruction term is amplified by $\lambda \sigma_i^{\lambda-1}$. As a result, the effective gradient ratio becomes $\rho_i^\lambda = \lambda \sigma_i^{\lambda-1} \rho_i$. As $\sigma_i \to 1$, $\rho_i^\lambda \to \lambda \rho_i$. A dimension with $\rho_i < \rho_{\text{crit}}$ (which would collapse under $\lambda = 1$) is protected once $\lambda > \rho_{\text{crit}}/\rho_i$.

Furthermore, a given $\lambda_i$ defines an attractor $\sigma_i$ at which point the amplification vanishes. The mechanism drives $\sigma$ above or below the stable point towards it; see Figure 2a and Appendix A. This behaviour leads to variance equalization across dimensions, which is the key mechanism of $\lambda$-VAE to counteract *polarization* and hence posterior collapse.




(a) Convergence to $\sigma^*$

(b) Optimal $\lambda_i^*(\sigma_i, \delta)$

Figure 2: **Variance equalization in $\lambda$-VAE.** *(a)* Convergence dynamics for $\lambda = 2$ ($\sigma^* = 0.5$): six dimensions starting at $\sigma_0 \in \{0.10, 0.20, 0.35, 0.65, 0.80, 0.90\}$ all converge to $\sigma^* = 0.5$. Trajectories are solutions of $d\sigma/dt = \sigma - \lambda\sigma^\lambda$ (the gradient correction term with fixed reconstruction signal), isolating the equalization mechanism. *(b)* Optimal scaling exponent $\lambda_i^*(\sigma_i, \delta)$ for $\delta \in \{2, 3, 5, 10\}$. Below $\sigma_c = \sqrt{1 - 1/\delta}$, $\lambda^* = 1$; above it, $\lambda^*$ increases with $\sigma_i$, allocating larger corrections to the most uncertain dimensions.

### 4.2 Information Gain

Under $\lambda$-scaling, using $H[Z] \approx H[Z_\lambda]$ (valid for moderate $\lambda$; Appendix A), we have the following information gain

$$I(X; Z_\lambda) - I(X; Z_1) = \frac{\lambda - 1}{2} \sum_{i=1}^{d} \mathbb{E}\big[-\log \sigma_i^2\big] \geq 0. \tag{12}$$

All terms are positive for $\sigma_i < 1$, giving a residual information gap $\Delta I_\lambda = \Delta I_1 \cdot (2 - \lambda)$. As such, for a given training snapshot, at $\lambda = 2$ the gap is eliminated, and for $\lambda \in (1, 2)$ it is partially closed. The information gain per dimension is proportional to $\mathbb{E}[-\log \sigma_i^2]$, and thus confident dimensions (small $\sigma_i$) yield larger information gains than dimensions with wide posterior.

### 4.3 Variance Equalization and the Equilibrium Target

Consider the gradient correction introduced by $\lambda$-scaling, except this time with respect to $\log \sigma_i$; written as $\nabla_i = (\lambda\sigma_i^\lambda - \sigma_i) \cdot g_i^\sigma$. Setting the gradient $\nabla_i = 0$, gives the equilibrium condition $\lambda\sigma_i^{\lambda-1} = 1$ which is solved by

$$\sigma_i^* = \lambda^{-1/(\lambda-1)}. \tag{13}$$

Full derivation and stability proof in Appendix A. This value is the same regardless of starting initial posterior variance. Hence, every dimension, whether near-collapsed or over-sharp, is drawn toward the same $\sigma^*$, provided they share the same $\lambda$. This is the content of *variance equalization*. In addition, the attractor defined by Eq. 13 is stable. That is, for $\sigma_i > \sigma^*$, $\nabla_i < 0$ (system pushes $\sigma_i$ down); for $\sigma_i < \sigma^*$, $\nabla_i > 0$ (system pushes $\sigma_i$ up). The correction is a bidirectional stable restoring force. Furthermore, the linearization rate is $1 - \lambda$ for $\lambda > 1$, so convergence is exponentially fast with rate $|\lambda - 1|$. Figure 2a illustrates the target $\sigma^*(\lambda)$ and convergence from multiple starting points.

Hence, the polarized $\sigma$ distribution that characterizes emergent collapse is corrected simultaneously across all dimensions. Similarly, the variance equalization leads to emergent distributed representation as opposed to sparse Hinton (1986); Bengio et al. (2013).

### 4.4 Optimal $\lambda^*$ per Dimension

Increasing $\lambda$ improves both gradient ratios and information gain but risks divergence from the prior, which is critical for random generation. The $\mathbb{KL}$ divergence from $\mathcal{N}(0, \mathbf{I})$ under the $\lambda$-scaling is

$$\mathbb{KL}(\lambda_i) = \frac{1}{2}\Big(\sigma_i^{2\lambda_i} - \sigma_i^2 - \log \sigma_i^{2\lambda_i} + \log \sigma_i^2\Big) > 0 \quad \text{for } \lambda_i > 1, \sigma_i < 1. \tag{14}$$

As a result, we define an objective that can be used to calculate optimal $\lambda$ such that the information gain is balanced with the $\mathbb{KL}$ divergence, as follows

$$\mathcal{J}(\lambda_i, \sigma_i) = (\lambda_i - 1)|\log \sigma_i| - \delta\, \mathbb{KL}(\lambda_i), \qquad \delta > 1. \tag{15}$$

The first term is the per-dimension gain from Eq. 12; the second term is used to penalize the additional $\mathbb{KL}$ divergence due to $\lambda_i$; it is weighted by a hyperparameter $\delta > 1$ controlling the tradeoff. The objective is strictly concave in $\lambda_i$ and has a unique maximum, which is given by the following proposition.

**Proposition 3** (Optimal $\lambda^*$)**.** *The objective $\mathcal{J}$ is strictly concave in $\lambda_i$. Its unconstrained maximum, subject to $\lambda_i \geq 1$, is:*

$$\lambda_i^* = \max\left(1,\ \frac{\log(1 - 1/\delta)}{2 \log \sigma_i}\right). \tag{16}$$

*Full derivation in Appendix A.*

The optimal solution provided by the above proposition has a direct interpretation as *optimal resource allocation*. The estimate $\lambda_i^* \propto 1/|\log \sigma_i|$, hence uncertain dimensions (large $\sigma_i$, small $|\log \sigma_i|$) receive larger corrections, while confident dimensions receive smaller ones or none at all.

Meanwhile, the hyperparameter $\delta$ plays a key role in determining the aggressiveness of correction regime. Importantly, it determines a critical threshold $\sigma_c$ as

$$\sigma_c := \sqrt{1 - 1/\delta} \tag{17}$$

For any $\sigma_i < \sigma_c$, $\lambda^* = 1$, the $\mathbb{KL}$ divergence outweighs the gains, hence the max in Eq. 16 passes one. The larger $\delta$, the less tolerant the system is of $\mathbb{KL}$ divergence. In the extreme case, as $\delta \to \infty$ then $\lambda_i^* \to 1$ for all dimensions, recovering the standard VAE. On the contrary, as $\delta \to 1^+$, the system grows more tolerant to disparity in $\mathbb{KL}$; approaching a deterministic autoencoder, see Figure 2b for this mechanism.

## 5 Related Works

Posterior collapse in VAEs has motivated research along three lines: modifying the $\mathbb{KL}$ term or constraining decoder capacity to address gradient imbalance; isolating and minimizing marginal mismatch to improve aggregate posterior alignment; and enriching the prior or posterior family. We discuss how each relates to $\lambda$-VAE.

In $\beta$-VAE Higgins et al. (2016); Burgess et al. (2018) ELBO is modified by weighting the $\mathbb{KL}$ term. Large $\beta > 1$ encourages disentanglement but can encourage the decoder not to use latent information early in training, leading to posterior collapse. In Fu et al. (2019), cyclical annealing of $\beta$ is introduced between near zero and its maximum value. In both cases, the training objective is modified globally, affecting both reconstruction and $\mathbb{KL}$ gradients symmetrically, while offering no per-dimension adaptivity. Free Bits Chen et al. (2016) imposes a per-dimension minimum $\mathbb{KL}$ floor, introducing discontinuous gradients but guaranteeing minimum information usage. $\delta$-VAE Razavi et al. (2019a) enforces a similar constraint differentiably, leaving the threshold per dimension to be tuned. None of the above approaches introduce adaptive control over information flow conditioned on local encoder confidence $\sigma_i$ as is done in $\lambda$-VAE. Regardless, since $\lambda$-VAE modifies the sampling step rather than the training objective, it is fully compatible with $\beta$-VAE, cyclical annealing, and other $\mathbb{KL}$-weighting schedules and can be composed freely.

Normalising flows Rezende & Mohamed (2015); Kingma et al. (2016) enrich the posterior family through invertible transformations, reducing the approximation gap at the cost of significant parameter overhead. Rosca et al. Rosca et al. (2018) show that even expressive posteriors do not guarantee good aggregate matching, consistent with our finding that the information gap and marginal mismatch are the same pathology.

Lagging inference He et al. (2019) updates the encoder more aggressively than the decoder to prevent early collapse, targeting gradient imbalance but not the information gap or marginal mismatch. Linear VAE analyses Lucas et al. (2019); Ichikawa & Hukushima (2024) formalize collapse as a phase transition, while Ichikawa et al. Ichikawa & Hukushima (2024) derive a closed-form collapse threshold as a function of data covariance and $\mathbb{KL}$ weight.

Alternatively, using the ELBO decomposition Hoffman & Johnson (2016), InfoVAE Zhao et al. (2019) introduces separate weights for $I_q(X;Z)$ and the marginal, enforcing aggregate matching via maximum mean discrepancy (MMD). Wasserstein autoencoders Tolstikhin et al. (2017) match the aggregate using optimal transport; adversarial approaches Makhzani et al. (2015); Mescheder et al. (2017) use discriminator training. All require integration over the dataset to estimate $q(z)$, adding compute cost and training complexity. VampPrior Tomczak & Welling (2018) learns a flexible prior as a mixture of posteriors, reducing structural mismatch but requiring Monte Carlo $\mathbb{KL}$ estimation. Our duality result Eq. 8 shows that $\lambda$-VAE addresses marginal mismatch implicitly through the information gap, without aggregate estimation.

## 6 Experiments

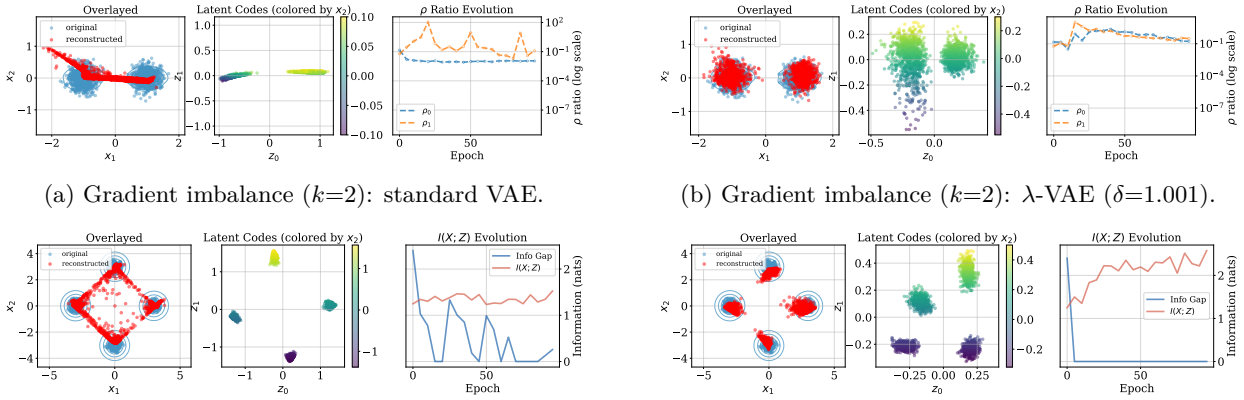
In this section we empirically validate $\lambda$-VAE on synthetic and real datasets, demonstrating its effect on posterior collapse, variance equalization, information flow, and generation quality.

### 6.1 Dataset and Setup

**Datasets:** We use a combination of synthetic and collected datasets. The synthetic dataset is drawn from a mixture of Gaussians $p(\mathbf{x}) = \sum_k \pi_k \mathcal{N}(\mathbf{x};\, \mu_k, \sigma_k I)$, with $k$=2 and $k$=4 components, generating two-dimensional data. For binary tasks we use Binarized MNIST Salakhutdinov & Murray (2008); Burda et al. (2015) and Binarized OMNIGLOT Lake et al. (2015); Burda et al. (2015); pixels are binarized stochastically as described in Burda et al. (2015). For RGB tasks we use CIFAR-10 Krizhevsky et al. (2009) (32×32, 10 classes) and CelebA-64 Liu et al. (2015) (face images center-cropped and resized to 64×64). All datasets use standard train–test splits and all results are reported on the test set.

**Models:** For synthetic and binarized datasets we use a 3-layer MLP with 300 hidden units per layer for both encoder and decoder. For RGB datasets we use an *asymmetric* ResNet: a weak encoder (base channels 16, 1 residual block per stage, 3 downsampling stages, channel multipliers $[1, 2, 4]$) paired with a strong decoder (base channels 64, 2 residual blocks per stage, 3 upsampling stages). The latent dimension is $K$=2 (synthetic), $K$=30 (MNIST and OMNIGLOT), $K$=512 (CIFAR-10), and $K$=1024 (CelebA-64). We use Bernoulli likelihood for binarized datasets and discretized logistic mixture (10 components) for RGB datasets. All likelihoods model pixels as statistically independent except in the cases where PixelCNN decoder Gulrajani et al. (2016) is used– we explicitly state it in such a case.

**Training:** Binary models are trained for 200 epochs with Adam Kingma & Ba (2014) (lr= 0.001). Models on Natural Images are trained for 400 epochs on 8 GPUs using Distributed Data Parallelism (DDP) with linear learning rate scaling (effective lr= $8\times10^{-4}$) Goyal et al. (2017), cosine decay to $\mathrm{lr}_{\min} = 10^{-5}$, and gradient clipping at 1.0, with 100 batch size. For optimal $\lambda^*$, the per-dimension values are updated every 5 epochs using an exponential moving average (EMA) of $\sigma$ (decay 0.9) and Eq. 16; a linear ramp of 150 epochs is applied to prevent premature compression before the encoder stabilizes. Meanwhile, the optimal $\lambda$ in both binary and synthetc dataset is updated every 2 epoch. We set $\delta$=1.01 for binary datasets and $\delta$=1.001 for image datasets, unless we are reporting abalation study on the hyperparameter. The baseline is an identical-architecture VAE with fixed $\lambda$=1 throughout training. Hence, the online difference between optimal $\lambda$-VAE and Standard VAE, is the per-dimension $\lambda$ update according to Eq. 16.




(a) Gradient imbalance ($k$=2): standard VAE.

(b) Gradient imbalance ($k$=2): $\lambda$-VAE ($\delta$=1.001).

(c) Information gap ($k$=4): standard VAE.

(d) Information gap ($k$=4): $\lambda$-VAE ($\delta$=1.001).

Figure 3: **Gradient imbalance and information gap on synthetic Gaussian mixtures.** Each panel tracks the gradient ratio $\rho_i$ and mutual information $I(X;Z)$ over training, alongside the latent space. *Top row* ($k$=2, clusters stretched along one axis), isolating gradient imbalance. In the standard VAE the reconstruction gradient $g_i^\sigma$ collapses to zero for the low-variance dimension before $\sigma_i \to 1$, confirming Proposition 1; the latent space shows complete information loss as well. $\lambda$-VAE amplifies the gradient by $\lambda\sigma_i^{\lambda-1}$, keeping $\rho_i$ elevated and the dimension active throughout training. *Bottom row* ($k$=4, equal inter-cluster separation along each axis): the standard VAE exhibits a fluctuating non-zero information gap, resulting in variation loss in both the latent and reconstruction spaces. Under $\lambda$-VAE with $\lambda^*$>2 for all dimensions the gap is driven to zero and the four-cluster structure is recovered, consistent with Section 4.2.

**Metrics:** We report NLL (nats) and bits per dimension BPD = NLL / $(D \ln 2)$ computed as the ELBO on the full test set[2]. Active units (AU) count dimensions with $A_{z_i} = \mathrm{Cov}_x(\mathbb{E}[z_i|x]) > 0.01$ Burda et al. (2015). For binary and image datasets we estimate $I(X;Z)$ via the SNR-based *information capacity* derived directly from the Gaussian channel model of Section 4.2 (Eq. 7). For synthetic data the true $I(X;Z)$ is bounded by the $K$=2 latent space and remains well below $\log N$ for any practical $N$; we therefore use the Monte Carlo mixture marginalisation estimator $\hat{I} = H(Z) - H(Z|X)$, where $H(Z)$ is approximated by the $N$-component posterior mixture and $H(Z|X)$ is evaluated in closed form. We report $C$ as **Capacity** in Tables 1 and 2, and the information gap $I(X;\phi) - C$ to track how much encoder capacity reaches the decoder. We additionally report **decoder capacity** $\sum_i C_i^{\mathrm{dec}}$, where

$$C_i^{\mathrm{dec}} = \sqrt{\mathbb{E}_x\left[\left(\frac{\partial \log p(x|z)}{\partial z_i}\right)^2\right]} \tag{18}$$

is the RMS decoder gradient per latent dimension (Tables 2 and 3); a dimension with $C_i^{\mathrm{dec}} \approx 0$ contributes nothing to the reconstruction regardless of $\sigma_i$.

### 6.2 Collapse Prevention, Information Recovery, and Reconstruction Quality

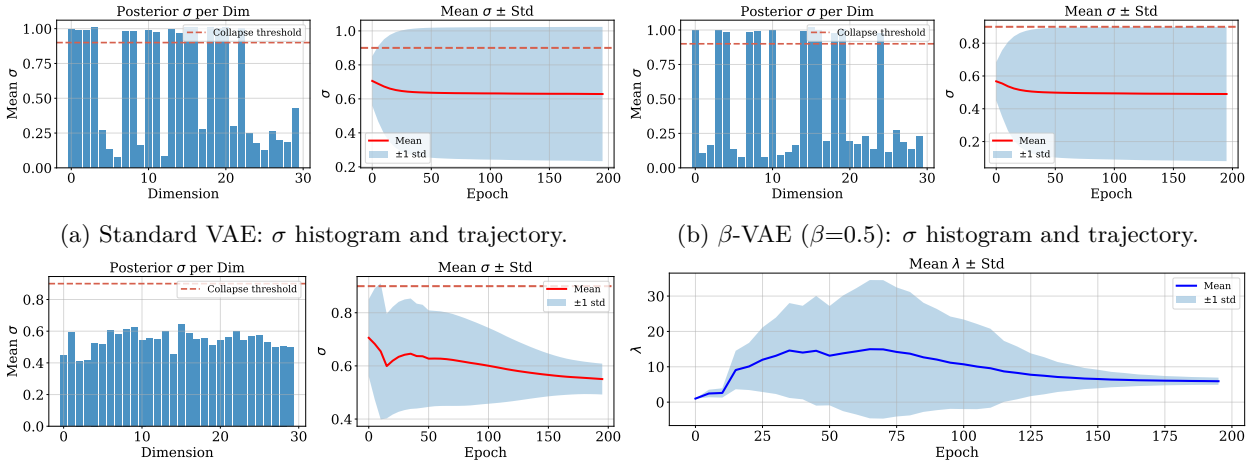
We validate $\lambda$-VAE along three axes, suppression of gradient imbalance, reduction of the information gap, and downstream reconstruction quality.

**Synthetic data:** We ran a standard VAE and $\lambda$-VAE on two synthetic $k$-component Gaussian mixtures with $K$=2 latent dimensions and $\delta$=1.001, designed to simulate the two causes of collapse independently (Figure 3). For $k$=2, the reconstruction gradient norm of the standard VAE's first dimension collapses to zero before $\sigma_1 \to 1$, matching Proposition 1; the latent space and reconstructions confirm complete information loss from that dimension. $\lambda$-VAE amplifies the gradient by $\lambda\sigma^{\lambda-1}$, keeping the dimension active

[2]IWAE Burda et al. (2015) is systematically biased for $\lambda$-VAE models: when $\sigma^{\lambda_d} \ll 1$ all importance samples cluster near $\mu$, collapsing the effective sample size and overestimating NLL. Hence, We report ELBO-BPD as the primary metric.

Table 1: **Comparison on binarized benchmarks (MNIST and Omniglot, $K$=30).** $\lambda$-VAE and the ablations shown ($\beta$-VAE, standard VAE) use a three-layer MLP encoder–decoder; most published baselines use deeper convolutional architectures, making these comparisons indicative rather than directly controlled. VampPrior is an exception, using a comparable MLP architecture. NLL and BPD measure reconstruction quality; AU counts active latent dimensions; Capacity is the SNR information capacity, measuring how much information reaches the decoder. Results for external baselines are taken from original papers.

| Dataset | Model | NLL↓ | BPD↓ | AU↑ | Capacity (nats)↑ |
|---|---|---|---|---|---|
| **MNIST** | Free Bits Kingma et al. (2016) | 79.10 | 0.145 | – | – |
| | AVB Mescheder et al. (2017) | 80.24 | 0.147 | – | – |
| | VampPrior Tomczak & Welling (2018) | 85.57 | 0.157 | – | – |
| | InfoVAE Zhao et al. (2019) | 80.76 | 0.148 | – | – |
| | $\beta$-VAE ($\beta$=0.5) Higgins et al. (2016) | 102.94 | 0.189 | 18 / 30 | 34.2 |
| | $\beta$-VAE ($\beta$=2.0) | 96.73 | 0.178 | 11 / 30 | 16.2 |
| | Standard VAE ($\lambda = 1.0$) | 97.17 | 0.178 | 14 / 30 | 23.2 |
| | **$\lambda$-VAE** ($\delta$=1.19) | 93.80 | 0.172 | 21 / 30 | 29.5 |
| | **$\lambda$-VAE** ($\delta$=1.10) | 94.44 | 0.173 | 23 / 30 | 37.2 |
| | **$\lambda$-VAE** ($\delta$=1.01) | **78.43** | **0.144** | **29 / 30** | **69.4** |
| **OMNIGLOT** | VampPrior | 104.75 | 0.192 | – | – |
| | $\beta$-VAE ($\beta$=0.5) | 131.65 | 0.242 | 21 / 30 | 38.4 |
| | $\beta$-VAE ($\beta$=2.0) | 119.86 | 0.220 | 13 / 30 | 15.0 |
| | Standard VAE ($\lambda = 1.0$) | 123.11 | 0.226 | 17 / 30 | 25.6 |
| | **$\lambda$-VAE** ($\delta$=1.19) | 111.14 | 0.204 | 29 / 30 | 32.4 |
| | **$\lambda$-VAE** ($\delta$=1.10) | 110.54 | 0.203 | 30 / 30 | 37.4 |
| | **$\lambda$-VAE** ($\delta$=1.01) | **103.03** | **0.189** | **30 / 30** | **67.7** |




(a) Standard VAE: $\sigma$ histogram and trajectory.

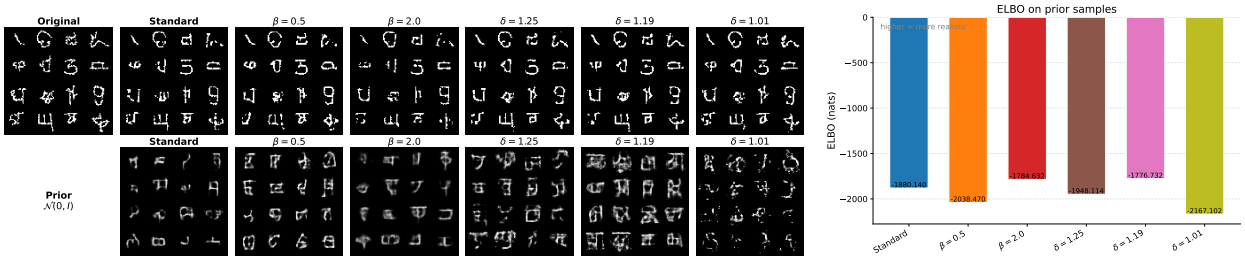
(b) $\beta$-VAE ($\beta$=0.5): $\sigma$ histogram and trajectory.

(c) $\lambda$-VAE ($\delta = 1.001$): variance equalization.

(d) $\bar{\lambda}^*$ trajectory (mean ± std over dimensions)

Figure 4: **Variance equalisation on Binary-MNIST ($K$=30).** Panels (a)–(c) each show the per-dimension $\sigma$ histogram alongside the mean ± std trajectory over training epochs. *(a)* Standard VAE: the distribution becomes bimodal; collapsed dimensions accumulate near $\sigma$=1 while active dimensions cluster at lower values. *(b)* $\beta$-VAE ($\beta$=0.5): the mode shifts downward but a tail near $\sigma$=1 persists. *(c)* $\lambda$-VAE ($\delta$=1.001): all dimensions converge to a tight band around $\sigma^*$; the across-dimension std collapses to near zero, realising variance equalization (Section 4.3). *(d)* Corresponding $\bar{\lambda}$ trajectory (mean ± std over dimensions): $\bar{\lambda}$ stabilises at $\approx$5.3, confirming joint $\sigma^*$–$\lambda^*$ convergence.

throughout training; compare Figure 3a vs 3b. For $k$=4, the standard VAE exhibits a fluctuating yet non-zero information gap that suppresses variation in both latent and data clusters, while introducing high reconstruction error per dimension leading to noisy data recovery. Under $\lambda$-VAE, however, since $\lambda^* > 2$ for




(a) Reconstruction vs Random sampling

(b) Prior samples ELBO

Figure 5: **Reconstruction vs Random sampling.** Reconstruction and prior sampling on Omniglot; figures best viewd zoomed. (a) Top: test reconstructions ($z$=$\mu$); bottom: samples from $z \sim \mathcal{N}(0, I)$ displayed as decoder probabilitie as pixel intensity for prob $> 0.5$ (greedy sampling of Bernoulli). Random samples are shared across models for the decoding. (b) ELBO on 256 prior samples per model variant. The figures show that $\delta$=1.19 retains reconstruction fidelity while producing coherent prior samples; $\delta$=1.01 achieves sharp reconstructions but loses generative quality as the aggregate posterior drifts from $\mathcal{N}(0, I)$, confirming that $\delta$ can be calibrated to close the information gap without over-compressing the latent code.

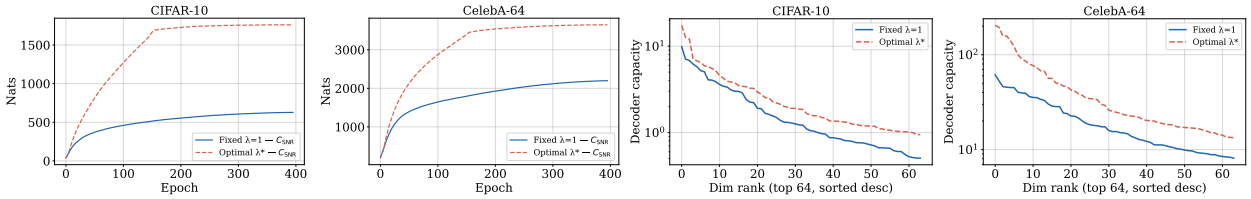
all dimensions the gap is driven to zero, consistent with Section 4.2, while both latent and reconstruction spaces preserve the four-cluster structure relatively well.

**Binary data:** On Binarized MNIST and Binarized OMNIGLOT ($K$=30, three-layer MLP), the standard VAE collapses 16 of 30 dimensions on MNIST (AU = 14) and 13 of 30 on OMNIGLOT (AU = 17). $\lambda$-VAE ($\delta$=1.01) reduces this to one on MNIST (AU = 29/30) and zero on OMNIGLOT (AU = 30/30), as reported in Table 1. In Figure 4, the mechanism is illustrated via $\sigma$ distributions. The standard VAE is bimodal with a mass near $\sigma$=1, while $\beta$-VAE shifts the distribution but does not eliminate the tail. The $\lambda$-VAE, however, concentrates all mass in a narrow band around $\sigma \approx 0.4$, realizing the variance equalization of Section 4.3. The average per-dimension estimate stabilizes at $\bar{\lambda} \approx 5.3$ on MNIST and $\bar{\lambda} \approx 2.5$ on OMNIGLOT at $\delta$=1.01, reflecting their different active-dimension $\sigma$ distributions; Section 6.3 confirms the theoretical predictive curve Eq. 16 aligns with emprically data.

The Capacity column in Table 1 separates two mechanisms for increasing information throughput, computed as Eq. 7. $\beta$-VAE ($\beta$=0.5) achieves comparable capacity to $\lambda$-VAE ($\delta$=1.1) on both datasets (34.2 vs. 37.2 nats on MNIST; 38.4 vs. 37.4 on OMNIGLOT), but does so by loosening the $\mathbb{KL}$ coefficient, inflating $\mathrm{Var}_x[\mu_d]$ across active dimensions at the expense of prior matching. The consequence is substantially worse BPD (0.242 vs. 0.203 on OMNIGLOT), while the aggregate posterior drifts from $\mathcal{N}(0, \mathbf{I})$. Provided a calibrated $\delta$ is choosen, $\lambda$-VAE instead reduces the effective noise $\sigma_d^{\lambda_d}$ per dimension without weakening the prior. At $\delta$=1.01, stronger per-dimension compression (avg. $\bar{\lambda}$≈5.3 on MNIST) drives Capacity to 69.4 nats (nearly 3× the standard VAE), while simultaneously achieving the best BPD. Meanwhile, $\delta = 1.19$ achievs a comparable AU, despite a significantly low infomration throughput – see Section 6.3 for discussion on effects of $\delta$.

**Natural images:** On RGB images, the fixed-$\lambda$ model on CIFAR-10 leaves 173 of 512 dimensions (AU=339), i.e., dimensions where the optimal schedule would assign $\lambda_d^* \gg 1$ but fixed $\lambda$=1 provides no correction. The SNR information capacity Eq. 7 (Table 2) sits at 628 nats. On the contrary under optimal $\lambda$, all 512 dimensions are driven below $\sigma$=0.9 and the effective noise $\sigma_d^{\lambda_d}$ is equalized across the full latent code; $C_{\mathrm{SNR}}$ grows to 1760 nats (2.8× improvement), as shown in Figure 6a. Decoder capacity rises in parallel from 169 to 361 nats (2.1×, Figure 6b) as a consequence of gradient rebalancing.

On CelebA-64 ($K$=1024), $C_{\mathrm{SNR}}$ grows from 2198 to 3646 nats (1.7×) and decoder capacity from 3283 to 6596 nats (2.0×). These gains translate directly into reconstruction quality. ELBO-BPD improves by +0.33 on CIFAR-10 (7.1% relative) and +0.24 on CelebA-64 (6.2% relative), see Table 2.




(a) SNR information capacity $C_{\text{SNR}}$ (b) Decoder capacity.

Figure 6: **SNR information capacity and decoder capacity over training.** (a) Under optimal $\lambda$, $C_{\text{SNR}}$ grows from 628 to 1760 nats on CIFAR-10 (2.8×) and from 2198 to 3646 nats on CelebA-64 (1.7×). (b) Decoder capacity rises in parallel: 169→361 nats (2.1×) on CIFAR-10 and 3283→6596 nats (2.0×) on CelebA-64, confirming that the increased encoder throughput is actively utilised by the decoder.

Table 2: **Controlled comparison on CIFAR-10 and CelebA-64.** Fixed and optimal $\lambda$ share the same asymmetric ResNet architecture. SNR capacity information throughput; decoder capacity (Figure 6b) measures the information each latent dimension actively contributes to the decoder output.

| **Dataset** | **Model** | **BPD ↓** | **SNR cap. ↑** | **Dec. cap. ↑** | $\bar{\lambda}$ | **AU ↑** |
|---|---|---|---|---|---|---|
| **CIFAR-10** | Vanilla VAE | 4.645 | 628 | 169 | 1.00 | 339 / 512 |
| | **λ-VAE** ($\delta$=1.001) | **4.314** | **1760** | **361** | 3.14 | **512 / 512** |
| **CelebA-64** | Vanilla VAE | 3.896 | 2198 | 3283 | 1.00 | 985 / 1024 |
| | **λ-VAE** ($\delta$=1.001) | **3.655** | **3646** | **6596** | 1.99 | **1024 / 1024** |

### 6.3 Optimal $\lambda$ and Its Sensitivity to $\delta$

**$\lambda^*$ estimate alignment with theory:** Figure 7 plots the per-dimension $(\sigma, \lambda^*)$ scatter for all latent dimensions in CIFAR-10 and CelebA-64 at epoch 400. Every dimension lies exactly on the theoretical curve defined in Propsotion 3, confirming that the EMA-based update converges to the analytic optimum derived in Appendix A. The fixed model's dimensions form a horizontal line at $\lambda$=1 across the full $\sigma$ range; the 173 dimensions with $\sigma > 0.9$ would receive $\lambda^* \gg 1$ under the optimal schedule, representing a systematic under-utilisation that fixed $\lambda$ cannot correct. The same alignment holds for CelebA-64 ($K$=1024; $\bar{\lambda}$=1.99, $\lambda_{\max}$=10.4 at convergence). The convergence of $\lambda^*$ and $\sigma^*$ is also shown in Figure 4 for Binary MNIST dataset.

**Effect of $\delta$:** The parameter $\delta$ derives the equilibrium $\sigma*$ via optimal estimate Eq. 16. It controls the agrresivness of compression (how close sample $z$ is to $\mu$). Smaller $\delta$ induces larger $\lambda^*$, compressing more dimensions more tightly. Table 1 illustrates this particularly on binary MNIST dataset; reducing $\delta$ from 1.19 to 1.01 increases AU from 21/30 to 29/30 and improves ELBO-BPD from 0.172 to 0.144. The AU ceiling, however, is reached at $\delta$=1.1 for OMNIGLOT.

As discussed in Section 4.4, the tradeoff here is increased marginal divergence for improved information flow, which consequently affects the quality of sampling from a prior. This can be calibrated as required. Figure 5 shows that $\delta$=1.19 navigates this tradeoff effectively – it improves reconstruction, BPD 0.204 vs. 0.226 for the standard VAE on OMNIGLOT, while achieving *best* prior sample quality than all compared methods (Standard VAE, $\beta$-VAE and different $\delta$ values) with −1776 ELBO, estimated on 256 prior samples. Meanwhile, $\delta$=1.01, while minimising BPD (0.189), substantially degrades generative quality (−2167 ELBO). Consequently, choosing the right $\delta$ leads to maximized net information gain Eq. 15, with little or no compromise on the generative equality.

### 6.4 Qualitative Comparision

In this section, we show the impact of increased information flow and minimized posterior collapse on reconstruction quality.

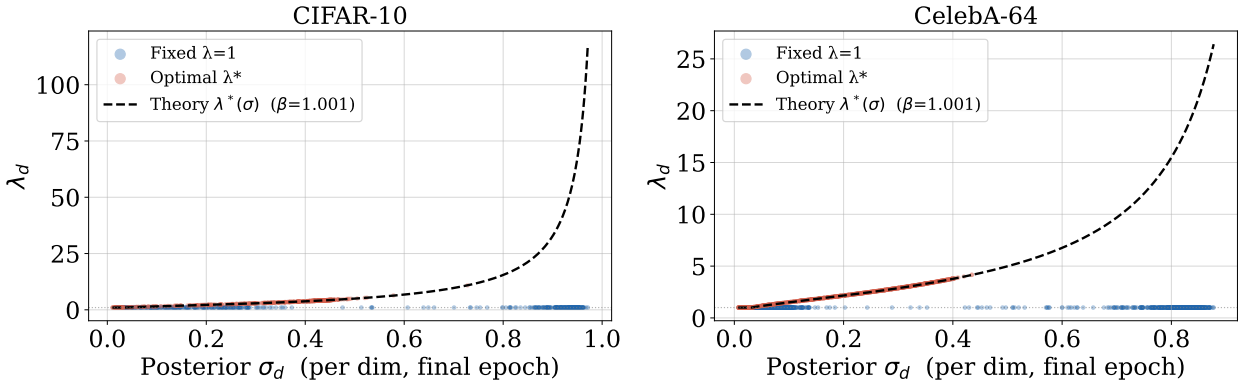
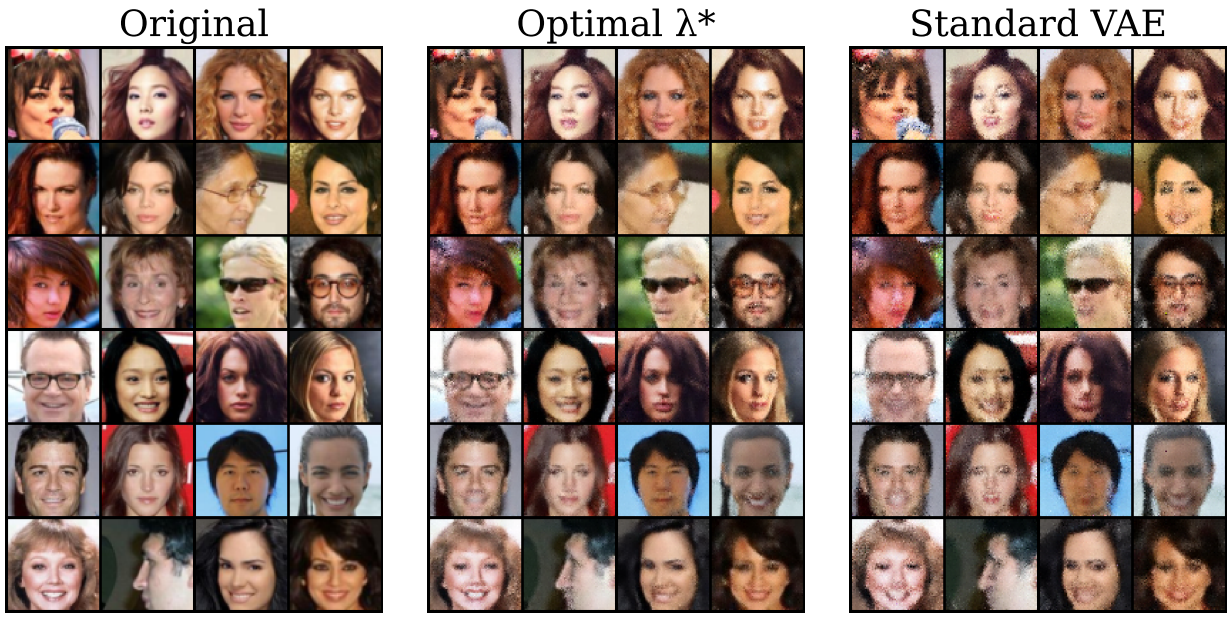



Figure 7: **Empirical alignment with the optimal $\lambda^*$ schedule.** All 512 CIFAR-10 and 1024 CelebA-64 dimensions lie exactly on the theoretical curve Eq. 16 at convergence (Pearson $r = 1.00$). Points near $\sigma \approx 0$ have $\lambda^* \to \lambda_{\max} = 20$ (clamped); points near $\sigma = 1$ are dimensions the vanilla VAE model leaves dead.

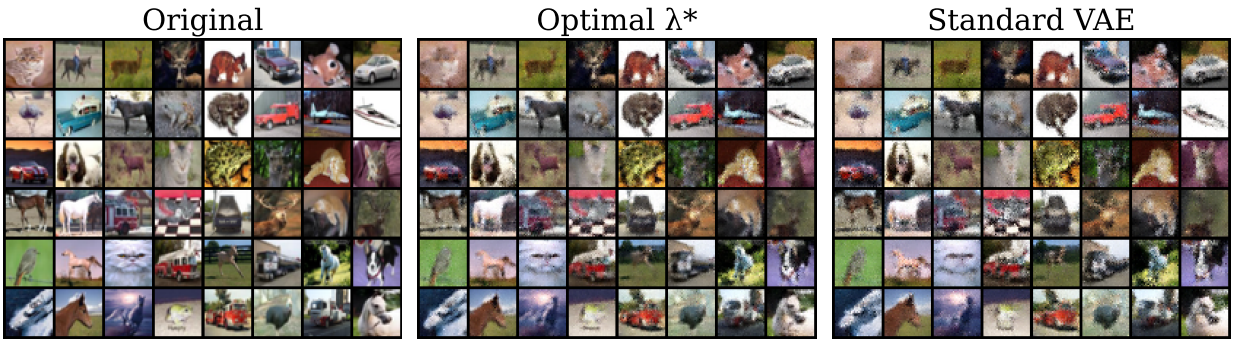
In Figure 8, we compare reconstructions between the standard and optimal-$\lambda$ VAE. $\lambda$-VAE produces sharper images with more coherent pixel values on both CIFAR-10 and CelebA-64. This qualitative improvement follows directly from the 2.8× gain in SNR information capacity and the 2.1× gain in decoder capacity shown in Figure 6 and Table 2. Reconstruction quality degradation due to posterior collapse, however, is much more prominently observed in a case where the modelling is primairly absorbed by the decoder. A sufficiently powerful decoder trained with standard VAE causes significant gradient imbalance that can render most of the latent code uninformative and wasted. This phenomenon is more apparent in autoregressive models like pixelCNN decoders Gulrajani et al. (2016) where the decoder is trained with teacher forcing and can quickly learn to ignore most of the latent codes, especially pixel level information when it is conditioned on the ground-truth of previous pixel. Hence, leading to a significant posterior collapse and poor sample quality during *inference*, despite a comparably good BPD during training. In contrast, the optimal $\lambda$ keeps $z_i$ close to $\mu_i$ with high SNR throughout training with reasonable variance (depending on $\delta$), ensuring the decoder cannot bypass the latent code regardless of decoder capacity. We show this contrast by using the same architecture as before only replacing the decoder with a PixelCNN decoder. Both models achieve comparable ELBO-BPD at evaluation (3.518 standard vs. 3.494 optimal; Table 3), yet 378 of 512 latent dimensions are effectively collapsed in the standard VAE versus none in the optimal $\lambda$ model. Decoder capacity further confirms this; the standard VAE's PixelCNN allocates only 12.1 nats to the latent code while $\lambda$-VAE allocates 75.0 nats (6.2× more), meaning the autoregressive decoder has learned to model CIFAR-10 almost entirely from its own context. The collapse is exposed when sampling from the model in both *greedy* mode (argmax at each pixel conditioned on $z$ and all previous pixels) and *ancestral* sampling ($p(x_i|x_{<i}, z)$ drawn sequentially per pixel), as shown in Figure 9. This is primarily due to preserved information flow under optimal $\lambda$-VAE even though teacher forcing was used in both cases for training.

## 7 Conclusion

Posterior collapse is commonly attributed to the $\mathbb{KL}$ term overwhelming reconstruction, but this framing misses the mechanism. The $\mathbb{KL}$ term does not suppress all dimensions uniformly – it drives *specific* dimensions toward $\sigma_i = 1$ through two coupled pathways. At $\sigma_i = 1$, both pathways converge: it is simultaneously a stable fixed point of the training dynamics where the reconstruction gradient vanishes (Proposition 1), and the regime in which injected noise overwhelms the encoded signal, reducing per-channel SNR and decoupling the decoder from the latent code (Proposition 2). $\lambda$-VAE corrects both by shifting the stable attractor from $\sigma_i = 1$ to $\sigma^* = \lambda^{-1/(\lambda-1)} < 1$, simultaneously restoring the reconstruction gradient and calibrating per-channel noise to signal strength. The per-dimension optimal $\lambda$ estimation (Proposition 3) follows from a net information gain objective that balances information throughput against marginal-prior divergence, with $\delta$ as the only hyperparameter. Empirically, $\lambda$-VAE consistently reduces collapsed dimensions, recovers information throughput, and improves reconstruction quality across all tested scales – synthetic, binarized,




(a) **CelebA-64**




(b) **CIFAR10**

Figure 8: **Qualitative Comparision On Recconstructions**. Optimal $\lambda$ produces sharper, more detailed reconstructions, consistent with the 2.1× capacity recovery in Table 2. Differences in quality are best observed zoomed in.

and large-scale RGB – with the PixelCNN experiment providing the sharpest demonstration with near-identical BPD (3.518 vs. 3.494) yet 6.2× more decoder capacity allocated to the latent code.

The broader implication on general training principles can be summaized as follows: a stochastic bottleneck trained end-to-end with a sufficiently powerful downstream module will be bypassed when the injected noise exceeds the signal capacity. This is not because the model cannot learn, but because bypassing is the lower-cost solution under a uniform noise budget. $\lambda$-VAE instantiates a targeted correction for Gaussian posteriors, but the same principle is expected to apply wherever the noise level at a stochastic layer is set globally rather than calibrated to per-channel signal strength. Three directions are left for future work. *Hierarchical VAEs:* collapse at early latent levels propagates through the hierarchy; per-level $\lambda^*$ schedules may provide targeted repair without the careful $\mathbb{KL}$ balancing that hierarchical training currently requires. *Adaptive $\delta$:* the current approach sets $\delta$ once per experiment; learning $\delta$ or scheduling it based on the evolving $\sigma$ distribution could recover the reconstruction, generation tradeoff automatically, and reduce per-dataset tuning. *Beyond VAEs:* the information gap–marginal mismatch duality (Eq. 8) provides a cheap, dataset-free proxy for aggregate

(a) Greedy decoding (argmax at each pixel step).

(b) Ancestral sampling from $p(x_i|x_{<i}, z)$.

Figure 9: **Reconstruction with PixelCNN decoder on CIFAR-10.** Each panel shows ground-truth (left), Standard VAE (middle), and $\lambda$-VAE (right) for the same test images. Despite near-identical ELBO-BPD (3.518 vs. 3.494, Table 3), the standard VAE has 378/512 collapsed latent dimensions and decoder capacity of only 12.1 nats: the PixelCNN has learned to model the image from pixel context alone and ignores $z$. This collapse is visible in both sampling modes (greedy and ancestral) where $\lambda$-VAE reconstructions retain input-specific detail even without teacher forcing.

Table 3: **PixelCNN decoder comparison on CIFAR-10.** BPD evaluated via the test-set ELBO. Decoder capacity sums per-dimension contributions from the training history (a proxy for how much the decoder relies on $z$).

| **Model** | **BPD**↓ | **AU**↑ | **Dec. cap. (nats)**↑ | $\bar{\lambda}$ |
|---|---|---|---|---|
| Standard VAE | 3.518 | 134 / 512 | 12.1 | 1.00 |
| $\lambda$-VAE ($\delta$=1.001) | **3.494** | **512 / 512** | **75.0** | 6.37 |

posterior quality that avoids the integration cost of InfoVAE or Wasserstein objectives; whether this proxy can guide training in discrete bottleneck models or diffusion-based posteriors is an open question.

## References

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.

Samuel R Bowman, Luke Vilnis, Oriol Vinyals, Andrew M Dai, Rafal Jozefowicz, and Samy Bengio. Generating sentences from a continuous space. *arXiv preprint arXiv:1511.06349*, 2015.

Yuri Burda, Roger Grosse, and Ruslan Salakhutdinov. Importance weighted autoencoders. *arXiv preprint arXiv:1509.00519*, 2015.

Christopher P Burgess, Irina Higgins, Arka Pal, Loic Matthey, Nick Watters, Guillaume Desjardins, and Alexander Lerchner. Understanding disentangling in beta-vae. *arXiv preprint arXiv:1804.03599*, 2018.

Xi Chen, Diederik P Kingma, Tim Salimans, Yan Duan, Prafulla Dhariwal, John Schulman, Ilya Sutskever, and Pieter Abbeel. Variational lossy autoencoder. *arXiv preprint arXiv:1611.02731*, 2016.

Thomas M Cover. *Elements of information theory*. John Wiley & Sons, 1999.

Hao Fu, Chunyuan Li, Xiaodong Liu, Jianfeng Gao, Asli Celikyilmaz, and Lawrence Carin. Cyclical annealing schedule: A simple approach to mitigating kl vanishing. *arXiv preprint arXiv:1903.10145*, 2019.

Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate, large minibatch sgd: Training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*, 2017.

Ishaan Gulrajani, Kundan Kumar, Faruk Ahmed, Adrien Ali Taiga, Francesco Visin, David Vazquez, and Aaron Courville. Pixelvae: A latent variable model for natural images. *arXiv preprint arXiv:1611.05013*, 2016.

Junxian He, Daniel Spokoyny, Graham Neubig, and Taylor Berg-Kirkpatrick. Lagging inference networks and posterior collapse in variational autoencoders. *arXiv preprint arXiv:1901.05534*, 2019.

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*, 2016.

Geoffrey E Hinton. Learning distributed representations of concepts. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 8, 1986.

Matthew D Hoffman and Matthew J Johnson. Elbo surgery: yet another way to carve up the variational evidence lower bound. In *Workshop in Advances in Approximate Bayesian Inference, NIPS*, volume 1, 2016.

Yuma Ichikawa and Koji Hukushima. Learning dynamics in linear vae: Posterior collapse threshold, superfluous latent space pitfalls, and speedup with kl annealing. In *International Conference on Artificial Intelligence and Statistics*, pp. 1936–1944. PMLR, 2024.

Michael I Jordan, Zoubin Ghahramani, Tommi S Jaakkola, and Lawrence K Saul. An introduction to variational methods for graphical models. *Machine learning*, 37:183–233, 1999.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

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

Durk P Kingma, Tim Salimans, Rafal Jozefowicz, Xi Chen, Ilya Sutskever, and Max Welling. Improved variational inference with inverse autoregressive flow. *Advances in neural information processing systems*, 29, 2016.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Brenden M Lake, Ruslan Salakhutdinov, and Joshua B Tenenbaum. Human-level concept learning through probabilistic program induction. *Science*, 350(6266):1332–1338, 2015.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pp. 3730–3738, 2015.

James Lucas, George Tucker, Roger B Grosse, and Mohammad Norouzi. Don't blame the elbo! a linear vae perspective on posterior collapse. *Advances in Neural Information Processing Systems*, 32, 2019.

Alireza Makhzani, Jonathon Shlens, Navdeep Jaitly, Ian Goodfellow, and Brendan Frey. Adversarial autoencoders. *arXiv preprint arXiv:1511.05644*, 2015.

Geoffrey J McLachlan and Thriyambakam Krishnan. *The EM algorithm and extensions*. John Wiley & Sons, 2007.

Lars Mescheder, Sebastian Nowozin, and Andreas Geiger. Adversarial variational bayes: Unifying variational autoencoders and generative adversarial networks. In *International conference on machine learning*, pp. 2391–2400. PMLR, 2017.

Thomas P Minka. Expectation propagation for approximate bayesian inference. *arXiv preprint arXiv:1301.2294*, 2013.

Radford M Neal and Geoffrey E Hinton. A view of the em algorithm that justifies incremental, sparse, and other variants. In *Learning in graphical models*, pp. 355–368. Springer, 1998.

Ali Razavi, Aäron van den Oord, Ben Poole, and Oriol Vinyals. Preventing posterior collapse with delta-vaes. *arXiv preprint arXiv:1901.03416*, 2019a.

Ali Razavi, Aaron Van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with vq-vae-2. *Advances in neural information processing systems*, 32, 2019b.

Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *International conference on machine learning*, pp. 1530–1538. PMLR, 2015.

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *International conference on machine learning*, pp. 1278–1286. PMLR, 2014.

Mihaela Rosca, Balaji Lakshminarayanan, and Shakir Mohamed. Distribution matching in variational inference. *arXiv preprint arXiv:1802.06847*, 2018.

Ruslan Salakhutdinov and Iain Murray. On the quantitative analysis of deep belief networks. In *Proceedings of the 25th international conference on Machine learning*, pp. 872–879, 2008.

Yuge Shi, Brooks Paige, Philip Torr, et al. Variational mixture-of-experts autoencoders for multi-modal deep generative models. *Advances in neural information processing systems*, 32, 2019.

Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. *arXiv preprint physics/0004057*, 2000.

Ilya Tolstikhin, Olivier Bousquet, Sylvain Gelly, and Bernhard Schoelkopf. Wasserstein auto-encoders. *arXiv preprint arXiv:1711.01558*, 2017.

Jakub Tomczak and Max Welling. Vae with a vampprior. In *International Conference on Artificial Intelligence and Statistics*, pp. 1214–1223. PMLR, 2018.

Shengjia Zhao, Jiaming Song, and Stefano Ermon. Infovae: Balancing learning and inference in variational autoencoders. In *Proceedings of the aaai conference on artificial intelligence*, volume 33, pp. 5885–5892, 2019.

# A Appendix

## A.1 Gradient Imbalance (Proposition 1)

The reconstruction gradient w.r.t. $\sigma_i$ is defined as

$$g_i^\sigma \;:=\; \frac{\partial R}{\partial \sigma_i} \;=\; \mathbb{E}_{p(\mathbf{x}),\,\epsilon_i}[\nabla_{z_i} \log p_\theta(\mathbf{x}|\mathbf{z}) \cdot \epsilon_i]\,, \tag{19}$$

while the $\mathbb{KL}$ gradient is $\partial \mathbb{KL}_i / \partial \sigma_i = \sigma_i - 1/\sigma_i$. Combining, the ELBO gradient is

$$\frac{\partial \mathcal{L}}{\partial \sigma_i} \;=\; g_i^\sigma + \frac{1}{\sigma_i} - \sigma_i. \tag{A.1}$$

From the above, we see the $\mathbb{KL}$ restoring force $1/\sigma_i - \sigma_i$ is positive for $\sigma_i < 1$, and negative for $\sigma_i > 1$. It drives $\sigma_i \to 1$ whenever it dominates the reconstruction gradient $g_i^\sigma$.

**Gradient ratio and fixed points:** Setting the ELBO gradient Eq. A.1 to zero gives the following equilibrium condition:

$$\sigma_i^* \;=\; \frac{-g_i^\sigma + \sqrt{(g_i^\sigma)^2 + 4}}{2} \;\in\; (0, 1), \tag{A.2}$$

When $g_i^\sigma = 0$, this gives the unique positive root $\sigma_i^* = 1$. Linearising around $\sigma_i = 1 + \delta$ and simplifying $\partial \mathcal{L} / \partial \sigma_i$ yields $1/\sigma_i - \sigma_i \approx -2\delta + O(\delta^2)$, so $\partial \mathcal{L} / \partial \sigma_i \approx -2\delta$. Thus, the collapse fixed point $\sigma_i^* = 1$ is a *locally*

*asymptotically stable point with linearisation rate* $-2$. For $g_i^\sigma > 0$, the gradient ratio is the necessary and sufficient collapse indicator stated in Proposition 1.

$$\rho_i \;=\; \frac{|g_i^\sigma|}{|1/\sigma_i - \sigma_i|}. \tag{A.3}$$

**Proof of sufficency:** Assume $\rho_i(t) < 1$ for all $t \geq T_0$. Then $|g_i^\sigma| < |1/\sigma_i - \sigma_i|$, so $\partial\mathcal{L}/\partial\sigma_i > 0$ whenever $\sigma_i < 1$. The sequence $\{\sigma_i(t)\}_{t\geq T_0}$ is monotonically non-decreasing and bounded above by 1. By the *Monotone Convergence Theorem* it converges to a limit $\sigma_i^\infty \leq 1$. If $\sigma_i^\infty < 1$, then $1/\sigma_i^\infty - \sigma_i^\infty > 0$, and Eq. A.1 requires $g_i^\sigma(\sigma_i^\infty) = 0$ at a fixed point, which forces $\sigma_i^\infty = 1$, which is a contradiction. Hence $\sigma_i^\infty = 1$. ■

**Proof of necessity:** Under the decay condition provided in the proposition, near $\sigma_i = 1$ the $\mathbb{KL}$ part is written as

$$\left|\frac{1}{\sigma_i} - \sigma_i\right| \;=\; \frac{(1-\sigma_i)(1+\sigma_i)}{\sigma_i} \;\sim\; 2(1-\sigma_i) \quad \text{as } \sigma_i \to 1^-, \tag{A.4}$$

while the decay bound $|g_i^\sigma| = O((1-\sigma_i)^\alpha)$ gives the ratio bounded by

$$\rho_i \;=\; \frac{O((1-\sigma_i)^\alpha)}{2(1-\sigma_i)} \;=\; O\big((1-\sigma_i)^{\alpha-1}\big)\,. \tag{A.5}$$

- $\alpha > 1$**:** $(1-\sigma_i)^{\alpha-1} \to 0$ as $\sigma_i \to 1^-$, so $\rho_i \to 0$. Collapse implies $\rho_i \to 0$. Hence, *The ratio is a* ***necessary and sufficient*** *collapse signal for dimension* $i$. ■
- $\alpha = 1$**:** $\rho_i \to g_0/2 > 0$ as $\sigma_i \to 1$: the ratio stays bounded away from zero regardless of whether collapse occurs, giving a potential *false negative*. The condition $\alpha > 1$ in the proposition is therefore not a restriction but a *regime identifier*; it characterises the decoders for which $\rho_i$ is an informative diagnostic.
- $\alpha < 1$**:** $\rho_i \to \infty$; the decoder resists collapse. No collapse occurs.

### A.2 Information Gap (Proposition 2)

We derive the gap $\Delta I = I(X;\phi) - I(X;Z) \geq 0$ under Gaussian posteriors and show it attenuates change in the encoder. By the *Data Processing Inequality* Cover (1999), we have

$$I(X;Z) \;\leq\; I(X;\phi) \qquad \Longrightarrow \qquad \Delta I \;=\; I(X;\phi) - I(X;Z) \;\geq\; 0. \tag{20}$$

Since $z_i = \mu_i + \sigma_i\epsilon_i$ with $\sigma_i > 0$ is not invertible (given $z_i = z$, we cannot recover $\epsilon_i$ without $\mu_i, \sigma_i$), the map is strictly lossy: $\Delta I > 0$ for all $\sigma_i > 0$. Equivalently, by the chain rule for mutual information, the gap satisfies

$$\Delta I \;=\; I(X;\phi \mid Z) \;\geq\; 0, \tag{A.6}$$

the conditional mutual information between $X$ and the encoder parameters given $Z$.

**Information gap under Gaussian posteriors:** Consider an approximate posterior estimate $q(z_i|x) = \mathcal{N}(\mu_i(x), \sigma_i^2(x))$, and its conditional entropy

$$H[Z_i|X] \;=\; \tfrac{1}{2}\log(2\pi e) + \tfrac{1}{2}\mathbb{E}_{p(x)}\big[\log\sigma_i^2(x)\big]\,. \tag{A.7}$$

Since $I(X;\phi) = H[\phi]$ is fixed by the encoder weights (encoder is deterministic), changes in the gap $\Delta I$ are driven by changes in $I(X;Z)$. Under approximate marginal matching $q(z) \approx \mathcal{N}(\mathbf{0}, \mathbf{I})$ (valid under optimized $\mathbb{KL}$ regularization), the accessible mutual information reduces to

$$I(X;Z) \;\approx\; \frac{1}{2}\sum_{i=1}^{d}\mathbb{E}_{p(x)}\big[-\log\sigma_i^2(x)\big] \;\geq\; 0. \tag{A.8}$$

The gap grows monotonically: $\partial(\Delta I)/\partial\sigma_i = 1/\sigma_i > 0$. At collapse ($\sigma_i \to 1$), dimension $i$'s gap reaches its maximum $I(X;\phi_i)$ while $I(X;Z_i) \approx -\frac{1}{2}\log\sigma_i^2 \to 0$, reflecting complete information loss.

**Derivation of the sensitivity equation:** From the definition, we have the identity

$$I(X;Z) \;=\; I(X;\phi) - \Delta I. \tag{A.9}$$

Treating $\Delta I$ as a function of $I(X;\phi)$ and differentiating Eq. A.9 along the training trajectory

$$\frac{\partial\, I(X;Z)}{\partial\, I(X;\phi)} \;=\; 1 - \frac{\partial\, \Delta I}{\partial\, I(X;\phi)}. \tag{A.10}$$

Since $\mathcal{R} = f(I(X;Z))$ for monotonically increasing $f$, the chain rule gives

$$\frac{\partial \mathcal{R}}{\partial\, I(X;\phi)} \;=\; \frac{\partial \mathcal{R}}{\partial\, I(X;Z)} \cdot \left(1 - \frac{\partial\, \Delta I}{\partial\, I(X;\phi)}\right). \tag{A.11}$$

This is equation Eq. 6 in the main text.

**Duality with marginal mismatch:** From the standard ELBO decomposition (Hoffman & Johnson, 2016),

$$\mathbb{KL}(q(z|x)\|p(z)) \;=\; I(X;Z) + \mathbb{KL}(q(z)\|p(z)). \tag{A.12}$$

Substituting Eq. A.9:

$$\mathbb{KL}(q(z)\|p(z)) \;=\; \mathbb{KL}(q(z|x)\|p(z)) - I(X;\phi) + \Delta I. \tag{A.13}$$

## Appendix B

### B.1 Equilibrium Variance Target $\sigma^* = \lambda^{-1/(\lambda-1)}$

Under $\lambda$-scaling, the reparameterization is $z_i = \mu_i + \sigma_i^\lambda \epsilon_i$. Differentiating the ELBO objective with respect to $\log \sigma_i$, the modification introduced by $\lambda$-scaling relative to standard sampling produces the following scaling term

$$\Delta\nabla_i := \frac{\partial \mathcal{L}_\lambda}{\partial \log \sigma_i} - \frac{\partial \mathcal{L}_1}{\partial \log \sigma_i} \;=\; \left(\lambda\, \sigma_i^\lambda - \sigma_i\right) \cdot g_i^\sigma, \tag{21}$$

where $g_i^\sigma := \mathbb{E}_{\epsilon_i}[\nabla_{z_i} \log p_\theta(x|z) \cdot \epsilon_i]$ is the reconstruction gradient. The derivation follows from the chain rule through $\sigma_i^\lambda$: $\partial(\sigma_i^\lambda \epsilon_i)/\partial \log \sigma_i = \lambda \sigma_i^\lambda$, compared to $\partial(\sigma_i \epsilon_i)/\partial \log \sigma_i = \sigma_i$ in the standard case.

The scaling term in Eq. 21 is zero iff $\lambda \sigma_i^\lambda = \sigma_i$, which gives the equilibrium condition for $\sigma_i$:

$$\begin{aligned} \lambda\, \sigma_i^{\lambda-1} \;&=\; 1 \\ \sigma_i^* \;&=\; \lambda^{-1/(\lambda-1)}. \end{aligned} \tag{22}$$

The above is valid for $\lambda > 1$.

**$\sigma^*$ is a stable attractor:** The scaling term Eq. 21 acts as a restoring force. For $g_i^\sigma > 0$, the following conditions hold

- $\sigma_i < \sigma^*$: then $\sigma_i^{\lambda-1} < 1/\lambda$, so $\lambda \sigma_i^\lambda < \sigma_i$, giving $\Delta\nabla_i < 0$. The gradient on $\log \sigma_i$ is negative: $\sigma_i$ is pushed *upward* toward $\sigma^*$.
- $\sigma_i > \sigma^*$: then $\sigma_i^{\lambda-1} > 1/\lambda$, so $\lambda \sigma_i^\lambda > \sigma_i$, giving $\Delta\nabla_i > 0$. The gradient on $\log \sigma_i$ is positive: $\sigma_i$ is pushed *downward* toward $\sigma^*$.

The force is bidirectional and vanishes only at $\sigma^*$: $\sigma^*$ is a stable fixed point. Every dimension, regardless of its initial value, is driven toward the *same* target. See Figure 2a and Figure 4 for illustartive and emprical data of this behaviour.

### B.2 Optimal Per-Dimension $\lambda_i^*$

**Objective:** For a single latent dimension $i$ with encoder variance $\sigma_i \in (0,1)$, the per-dimension net information gain objective (Eq. 15 in the main text) is:

$$\mathcal{J}(\lambda_i, \sigma_i) \;=\; (\lambda_i - 1)\,|\log \sigma_i| \;-\; \delta\,\mathbb{KL}(\lambda_i), \qquad \delta > 1, \tag{23}$$

where $|\log \sigma_i| = -\log \sigma_i > 0$ for $\sigma_i \in (0,1)$, and the marginal $\mathbb{KL}$ cost (Eq. 14 in the main text) is:

$$\mathbb{KL}(\lambda_i) \;=\; \frac{1}{2}\Big(\sigma_i^{2\lambda_i} - \sigma_i^2 - \log \sigma_i^{2\lambda_i} + \log \sigma_i^2\Big). \tag{24}$$

**Derivative of $\mathcal{J}$ with respect to $\lambda_i$:** The information gain term contributes $|\log \sigma_i|$, which is constant in $\lambda_i$. For the marginal cost, the only $\lambda_i$-dependent terms in Eq. 24 are $\sigma_i^{2\lambda_i}$ and $\log \sigma_i^{2\lambda_i} = 2\lambda_i \log \sigma_i$. Hence,

$$\frac{\partial \mathbb{KL}}{\partial \lambda_i} \;=\; \frac{1}{2}\Big(2\sigma_i^{2\lambda_i} \log \sigma_i - 2\log \sigma_i\Big) \;=\; \log \sigma_i \Big(\sigma_i^{2\lambda_i} - 1\Big). \tag{25}$$

For $\sigma_i \in (0,1)$: $\log \sigma_i < 0$ and $\sigma_i^{2\lambda_i} < 1$, so $\sigma_i^{2\lambda_i} - 1 < 0$. Therefore $\partial \mathbb{KL}/\partial \lambda_i = \log \sigma_i(\sigma_i^{2\lambda_i} - 1) > 0$: the marginal cost is strictly increasing in $\lambda_i$. The full derivative of the objective is

$$\frac{\partial \mathcal{J}}{\partial \lambda_i} \;=\; |\log \sigma_i| - \delta \log \sigma_i \Big(\sigma_i^{2\lambda_i} - 1\Big). \tag{26}$$

**Stationary-point condition:** Setting Eq. 26 to zero,

$$|\log \sigma_i| \;=\; \delta \log \sigma_i \Big(\sigma_i^{2\lambda_i} - 1\Big). \tag{27}$$

Substituting $|\log \sigma_i| = -\log \sigma_i$ (since $\sigma_i < 1$):

$$-\log \sigma_i \;=\; \delta \log \sigma_i \Big(\sigma_i^{2\lambda_i} - 1\Big). \tag{28}$$

Dividing both sides by $\log \sigma_i \neq 0$:

$$-1 \;=\; \delta\Big(\sigma_i^{2\lambda_i} - 1\Big). \tag{29}$$

Rearranging,

$$\sigma_i^{2\lambda_i} \;=\; 1 - \frac{1}{\delta}. \tag{30}$$

**Solving for $\lambda_i^*$:** Taking the natural logarithm of both sides of Eq. 30:

$$2\lambda_i^* \,\log \sigma_i \;=\; \log\Big(1 - \frac{1}{\delta}\Big). \tag{31}$$

Solving for $\lambda_i^*$:

$$\lambda_i^* \;=\; \frac{\log(1 - 1/\delta)}{2\log \sigma_i}. \tag{32}$$

For $\sigma_i \in (0,1)$: $\log \sigma_i < 0$. For $\delta > 1$: $1 - 1/\delta \in (0,1)$, so $\log(1 - 1/\delta) < 0$. Therefore $\lambda_i^* > 0$. Enforcing the constraint $\lambda_i \geq 1$:

$$\boxed{\lambda_i^* \;=\; \max\Big(1,\; \frac{\log(1 - 1/\delta)}{2\log \sigma_i}\Big).} \tag{33}$$

**The stationary point is a global maximum:** The second derivative of Eq. 23:

$$\frac{\partial^2 \mathcal{J}}{\partial \lambda_i^2} \;=\; -\delta\, \frac{\partial^2 \mathbb{KL}}{\partial \lambda_i^2}. \tag{34}$$

From Eq. 25:

$$\frac{\partial^2 \mathbb{KL}}{\partial \lambda_i^2} \;=\; \frac{\partial}{\partial \lambda_i}\Big[\log \sigma_i \Big(\sigma_i^{2\lambda_i} - 1\Big)\Big] \;=\; 2\,(\log \sigma_i)^2\, \sigma_i^{2\lambda_i} \;>\; 0. \tag{35}$$

Therefore:

$$\frac{\partial^2 \mathcal{J}}{\partial \lambda_i^2} \;=\; -2\delta\,(\log \sigma_i)^2\, \sigma_i^{2\lambda_i} \;<\; 0. \tag{36}$$

The objective is strictly concave in $\lambda_i$, confirming that the stationary point Eq. 33 is a global maximum. Hence, maximizing the net information gain for a given $\delta$ hyperparameter.