OLVAE: One-Loss Variational Autoencoder for Generative Modeling without a Divergence Term

Joud Mesfer A. Alharthi*
KAUST Academy
KAUST, Thuwal, Saudi Arabia
joud2004as@gmail.com

Bushra Yasir Almonla*
KAUST Academy
KAUST, Thuwal, Saudi Arabia
bushramonla2004@gmail.com

Shahad Abdullah Aldeiji*
KAUST Academy
KAUST, Thuwal, Saudi Arabia
shahad.adj@gmail.com

Jad Mounayer
PIMM Laboratory
ENSAM, Paris, France
jad.mounayer@outlook.com

Abstract

We present OLVAE (One-Loss Variational Autoencoder), a generative autoencoder that reconstructs and generates images without a KL-divergence term. Standard VAEs balance reconstruction against a KL term through a weight β that must be re-tuned for each dataset, metric, and data regime. OLVAE removes this trade-off. It compresses the encoder output with a truncated SVD and applies a signed, identity-centred multiplicative Gaussian transformation to the retained coefficients, giving a stochastic latent space that can be sampled for generation while training with a single reconstruction loss. Across synthetic Gaussian data, Frey Faces, dSprites, and Shapes3D, OLVAE is competitive with a VAE, and better in several settings, without any β selection. It attains FID 28.40 on Frey Faces (comparable to the best-tuned VAE at 27.95, $\beta = 10^{-3}$, but with better reconstruction), the highest validity (0.953) and lowest JS divergence (0.139) on dSprites, and FID 47.91 on Shapes3D, compared with 132.61 ± 161.38 for the best tested VAE, while showing substantially lower variability. Its practical advantage is particularly clear in low-data settings, where the preferred VAE β changes with training-set size, whereas OLVAE requires no β selection. The main current limitation is a dependence on mini-batch size.

Code: github.com/bshra0/OLVAE-One-Loss-VAE

1 Introduction

A central goal of generative modelling is to approximate the probability distribution underlying the observed data so that new samples can be drawn from the learned model for synthesis and simulation [Kingma and Welling]. Classical probabilistic models approach this goal under different structural assumptions. Gaussian mixture models represent a density using a mixture of Gaussian components [Melnykov and Maitra, 2010], while hidden Markov models describe sequential observations through latent states governed by Markovian transitions [Rabiner, 1989]. In richer Bayesian models, exact posterior inference may be intractable, motivating approximate methods such as Markov chain Monte Carlo and variational inference [Blei et al., 2017].

Variational autoencoders (VAEs) provide a scalable framework for learning and approximate inference in deep latent-variable models [Kingma and Welling, 2014, Kingma and Welling]. Rather than

*These three authors contributed equally.

Preprint.

mapping each observation to a single fixed code, the probabilistic encoder defines an approximate posterior $q_\phi(z \mid x)$, while the decoder defines the conditional distribution $p_\theta(x \mid z)$. The β-weighted training objective used in this study is

$$\mathcal{L}_{\mathrm{VAE}}(x) = \mathcal{L}_{\mathrm{rec}}(x) + \beta D_{\mathrm{KL}}\left(q_\phi(z \mid x) \,\|\, p(z)\right), \tag{1}$$

where $\mathcal{L}_{\mathrm{rec}}$ denotes the reconstruction loss, β controls the relative contribution of the KL-divergence term, and the prior is $p(z) = \mathcal{N}(0, I)$. Setting $\beta = 1$ gives the unit KL weight used for the VAE baseline in this study, while varying β changes the relative contribution of the KL term, following the β-VAE formulation [Higgins et al., 2017]. VAEs have been applied across broad scientific domains, including chemistry [Gómez-Bombarelli et al., 2018], physics [Wetzel, 2017], and biology [Lopez et al., 2018].

The reconstruction and KL terms introduce a trade-off whose behaviour depends on their relative weighting [Higgins et al., 2017]. Against this background, two limitations are frequently reported in VAEs. One is over-smoothed or blurry output [Larsen et al., 2016, Czolbe et al., 2020, Tomczak and Welling, 2018]. The other is posterior collapse: if KL regularisation dominates, the approximate posterior may approach the prior and the latent variables may become uninformative [Bowman et al., 2016, He et al., 2019, Razavi et al., 2019, Lucas et al., 2019].

Increasing β in Equation (1) places greater relative emphasis on prior matching and may suppress latent information, whereas decreasing it emphasises reconstruction but may weaken prior matching. The effective balance is data- and implementation-dependent, so a value that works well in one setting may not transfer reliably to another. Extensive work has therefore sought to control this trade-off through reweighting, thresholding, scheduling, constraints on the variational family, modified inference optimisation, and alternative distribution-matching objectives [Higgins et al., 2017, Kingma et al., 2016, Fu et al., 2019, He et al., 2019, Razavi et al., 2019, Zhao et al., 2019, Tolstikhin et al., 2018]. Although these approaches differ, they all retain an explicit latent-space regularisation or distribution-matching mechanism during training.

Data availability provides a useful lens through which to examine this trade-off. When the number of observations is large relative to the complexity of the data distribution, the training data may cover the relevant factors of variation more densely, and a VAE may perform well with a relatively small KL weight. In low-data regimes, however, the latent representation is learned from fewer observations, making prior matching more consequential for organising regions of the latent space that are not densely supported by encoded training examples. The KL weight must therefore be large enough to encourage prior matching without suppressing informative latent codes. Consequently, the choice of β may become particularly important in data-limited settings. This motivates the central question of this work: can a generative latent space be regularised without introducing a reconstruction–KL trade-off?

Rather than modifying the weighting or scheduling of latent regularisation, rank-reduction autoencoders (RRAEs) [Mounayer et al., 2024] take a different route. In their strong formulation, they use SVD within the network to retain the dominant components of the latent representation, allowing training with reconstruction loss alone. As originally proposed, however, RRAEs are deterministic and not designed as generative models. Variational rank-reduction autoencoders (VRRAEs) make this framework probabilistic, but do so by reintroducing a KL-divergence term. The training objective therefore once again balances reconstruction against KL regularisation [Mounayer et al., 2025].

Building on this framework, we introduce structured stochasticity directly into the latent representation while retaining a single reconstruction objective with no divergence term. We therefore propose the *One-Loss Variational Autoencoder* (OLVAE), a new generative autoencoder built on this framework. After the SVD step, stochasticity is introduced through a specifically designed multiplicative noise matrix. This design enables generative sampling while training the model using reconstruction loss alone, without KL divergence. It therefore removes the coefficient that balances reconstruction against KL regularisation, simplifying the training objective. Our experiments reveal a regime-dependent pattern. VAEs can perform well with relatively small β values in data-richer settings, whereas reducing the amount of training data makes performance more sensitive to the reconstruction–KL balance, with the best-performing tested value of β changing across training-set sizes. We also observe posterior collapse in isolated VAE runs, indicating seed-dependent variability rather than uniform failure across runs. Our method removes this reconstruction–KL balancing parameter altogether and remains stable across the evaluated low-data settings. We evaluate it on controlled synthetic Gaussian data of known intrinsic dimension, the dSprites [Matthey et al., 2017] and 3D Shapes [Burgess and

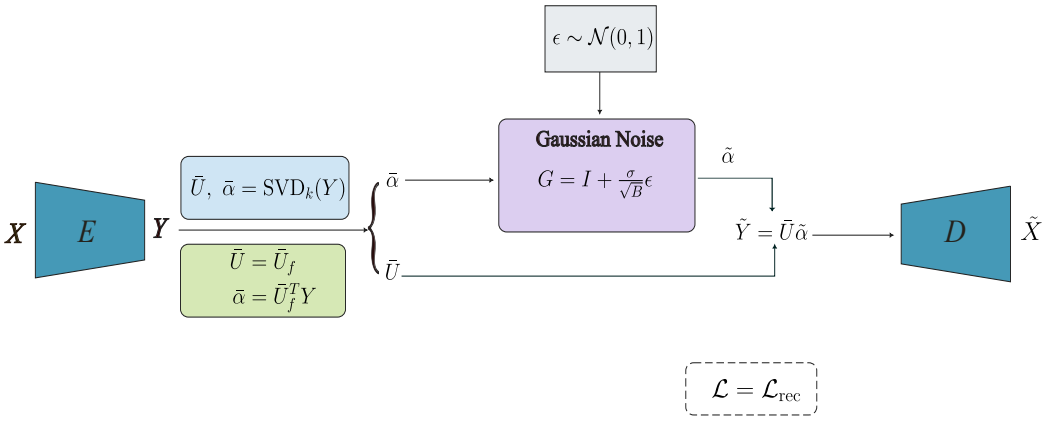


Figure 1: Overview of OLVAE. The encoder output is rank-reduced using truncated SVD, after which the retained coefficient matrix is right-multiplied by G_B. The resulting latent representation is then recomposed and passed to the decoder. Training minimizes mean-squared reconstruction error only.

Kim, 2018] benchmarks with known generative factors, and real face images from the Frey Face dataset [Frey, n.d.].

Our contributions are as follows.

- A generative autoencoder trained with a single reconstruction objective, with no KL or auxiliary divergence loss and hence no reconstruction–KL balancing coefficient to tune.
- A stochastic, data-adaptive latent mechanism that applies a signed, identity-centred multiplicative Gaussian perturbation to truncated SVD coefficients, so that sampled latent representations remain within the learned subspace.
- An evaluation on controlled synthetic Gaussian data, the dSprites and 3D Shapes benchmarks, and real Frey Face images, together with a VAE β-sensitivity study across the tested values and datasets.

2 Methodology

OLVAE combines truncated-SVD-based low-rank latent regularization with right-multiplication of the retained SVD coefficient matrix by an identity-centred Gaussian matrix (Fig. 1). The model is optimized using reconstruction error alone, while frozen aggregate SVD factors are used for inference and generation.

2.1 Model and rank-reduced latent representation

Let $X = [x_1, \ldots, x_B] \in \mathbb{R}^{d_x \times B}$ denote a mini-batch of B vectorized samples arranged as columns. In the implementation, the spatial tensor form is retained at the network input; vectorization is used here only for notation. The encoder and decoder are

$$E_\phi : \mathbb{R}^{d_x} \to \mathbb{R}^{d_z}, \qquad D_\theta : \mathbb{R}^{d_z} \to \mathbb{R}^{d_x}, \tag{2}$$

and are applied column-wise. The encoder produces the latent matrix

$$Y = E_\phi(X) \in \mathbb{R}^{d_z \times B}. \tag{3}$$

For a matrix $Y \in \mathbb{R}^{d_z \times B}$, the thin SVD can contain at most $q = \min(d_z, B)$ singular components. Let k be the prescribed bottleneck rank. The number of components actually retained is therefore $r = \min(k, q) = \min(k, d_z, B)$, and the thin SVD of Y is truncated accordingly:

$$Y = U\Sigma V^\top, \qquad \bar{Y} = U_r \Sigma_r V_r^\top = U_r A_r, \qquad A_r = \Sigma_r V_r^\top, \tag{4}$$

where $U_r \in \mathbb{R}^{d_z \times r}$, $\Sigma_r \in \mathbb{R}^{r \times r}$, $V_r^\top \in \mathbb{R}^{r \times B}$, and $A_r \in \mathbb{R}^{r \times B}$. Although $\bar{Y}$ remains in the d_z-dimensional decoder input space, it lies in $\mathrm{span}(U_r)$ and therefore has at most r effective degrees of freedom.

2.2 Multiplicative Gaussian transformation after SVD

After the truncated SVD, identity-centred multiplicative Gaussian noise is applied by right-multiplying the retained coefficient matrix by

$$G_B = I_B + \frac{\sigma}{\sqrt{B}}\Xi, \qquad \Xi_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1), \tag{5}$$

giving

$$A_r^G = A_r G_B, \qquad Z = U_r A_r^G = U_r \Sigma_r V_r^\top G_B. \tag{6}$$

Thus, G_B mixes the retained sample coefficients within the current mini-batch using signed Gaussian weights centred at the identity. The main configuration uses the signed-noise form in Eq. (5) with $\sigma = 0.1$; alternative definitions of G_B are considered only as ablations.

Expanding the product gives

$$A_r^G = A_r + \frac{\sigma}{\sqrt{B}} A_r \Xi, \tag{7}$$

so the transformation is multiplicative with respect to A_r and centred at the clean coefficients. To make the induced Gaussian structure explicit, let $q_i = (V_r^\top)_{:i} \in \mathbb{R}^r$ denote the retained coordinates of sample i (the i-th column of $V_r^\top$), let $a_i = \Sigma_r q_i$ be the corresponding coefficient column of A_r, and let $q_i^G = (V_r^\top G_B)_{:i}$ and $a_i^G = (A_r G_B)_{:i}$ be their transformed counterparts. Because Ξ has i.i.d. standard-normal entries and the rows of $V_r^\top$ are orthonormal ($V_r^\top V_r = I_r$),

$$q_i^G \mid V_r \sim \mathcal{N}\left(q_i, \frac{\sigma^2}{B} I_r\right), \qquad a_i^G \mid V_r, \Sigma_r \sim \mathcal{N}\left(a_i, \frac{\sigma^2}{B}\Sigma_r^2\right). \tag{8}$$

Therefore, conditioned on the retained SVD factors of the current mini-batch, G_B induces a Gaussian perturbation of each retained coefficient vector, whose conditional mean equals its clean value. This conditional statement does not imply that the unconditional (marginal) latent distribution is Gaussian or standard normal.

2.3 Forward path and reconstruction-only objective

A fresh matrix Ξ is sampled at every training forward pass. The complete forward path is

$$\widehat{X} = D_\theta\left(U_r \Sigma_r V_r^\top \left[I_B + \frac{\sigma}{\sqrt{B}}\Xi\right]\right). \tag{9}$$

The encoder and decoder parameters are optimized using only mean-squared reconstruction error:

$$\mathcal{L}_{\text{total}}(X, \Xi) = \mathcal{L}_{\text{rec}}(X, \Xi) = \frac{1}{B d_x}\left\|X - \widehat{X}\right\|_F^2. \tag{10}$$

No KL-divergence term or auxiliary distribution-matching loss is used. Automatic differentiation propagates gradients through the truncated SVD and the complete stochastic forward path.

2.4 Frozen-basis inference and generation

The batch-specific SVD basis is replaced at inference by a fixed aggregate basis. During training, the clean factors (U_b, Σ_b) and batch size B_b from each of the most recent $M \leq 20$ mini-batches are retained. They are aggregated as

$$W = [U_1\Sigma_1, \ldots, U_M\Sigma_M] = U_f \Sigma_f V_f^\top, \qquad N_{\text{rep}} = \sum_{b=1}^{M} B_b, \tag{11}$$

and the leading $r_f = \min(k, d_z, \sum_{b=1}^{M} r_b)$ components are retained, where $r_b = \min(k, d_z, B_b)$. Here, N_{rep} counts retained sample occurrences rather than necessarily distinct training samples. For an unseen input x, deterministic reconstruction is

$$\widehat{x} = D_\theta\big(U_f U_f^\top E_\phi(x)\big)\,. \tag{12}$$

No decoder fine-tuning is performed after the frozen basis is constructed.

Generation uses neither the encoder nor G_B. Instead, with $\eta \sim \mathcal{N}(0, I_{r_f})$, the implemented sampler is

$$y_{\text{new}} = \sqrt{\frac{1+\sigma^2}{N_{\text{rep}}}}\, U_f \Sigma_f \eta, \qquad x_{\text{new}} = D_\theta(y_{\text{new}}). \tag{13}$$

This sampling rule uses the directions and singular values of the aggregate frozen basis, with scaling determined by N_{rep} and the training noise level σ. It defines the implemented generation procedure without claiming that the complete marginal latent distribution is Gaussian.

3 Results

In this section, we evaluate OLVAE on a synthetic Gaussian dataset and three real-world datasets: Frey Faces, dSprites, and Shapes3D. We compare OLVAE against a Variational Autoencoder (VAE) and a deterministic Rank Reduction Autoencoder (RRAE). For training, OLVAE, VAE, and RRAE share the same base architecture (i.e., encoder and decoder) within each experiment, unless otherwise stated. The standard VAE baseline uses $\beta = 1$. We additionally perform a β-sensitivity study to assess the effect of KL weighting and report the best tested β where applicable.

3.1 Synthetic Data

We first evaluate the models on a synthetic dataset consisting of 64×64 images of a Gaussian bump with fixed magnitude and spread, displaced along both spatial axes. The dataset therefore has an intrinsic dimension of two. Training centres are placed on an equidistant grid, while test centres are sampled uniformly within the same range. Accordingly, the VAE uses a two-dimensional latent space, while RRAE and OLVAE retain two components. Generation quality is evaluated by fitting a Gaussian to each generated sample and reporting the relative error, together with the spread of the generated centres to assess sample diversity.

Higher-data regime (β sensitivity at $N_{\text{tr}} = 400$)

At $N_{\text{tr}} = 400$, the VAE shows strong sensitivity to β. Small values achieve low reconstruction and generation errors, while performance degrades substantially as β increases. The best tested VAE configurations achieve lower reconstruction and generation errors than OLVAE, although the best tested β differs between the two metrics. OLVAE maintains meaningful reconstruction and generation without KL regularization or β tuning. The full β sweep is reported in Table 1.

Table 1: Synthetic Gaussian results for $N_{\mathrm{tr}} = 400$ and $N_{\mathrm{te}} = 100$. The VAE is evaluated across the tested β values and compared with OLVAE. Values are reported as mean $\pm$ sample standard deviation over five seeds.

Model	Test recon. MSE ↓	Gen. error (%) ↓
VAE, $\beta = 0^{\dagger}$	0.00407 ± 0.00306	1.75 ± 0.91
VAE, $\beta = 10^{-6}$	0.00399 ± 0.00341	$\mathbf{1.46 \pm 0.17}$
VAE, $\beta = 10^{-5}$	0.00413 ± 0.00177	1.89 ± 0.62
VAE, $\beta = 10^{-4}$	$\mathbf{0.00339 \pm 0.00125}$	2.35 ± 0.90
VAE, $\beta = 10^{-3}$	0.00433 ± 0.00089	2.40 ± 0.51
VAE, $\beta = 10^{-2}$	0.01022 ± 0.00398	3.40 ± 0.69
VAE, $\beta = 10^{-1}$	0.02857 ± 0.00520	8.48 ± 0.56
VAE, $\beta = 1$	0.80278 ± 0.02061	71.43 ± 0.82
VAE, $\beta = 10$	0.80292 ± 0.02062	71.30 ± 0.89
VAE, $\beta = 64$	0.80282 ± 0.02050	71.53 ± 0.85
OLVAE (no β)	0.01048 ± 0.00361	5.49 ± 1.83

$^{\dagger}\beta = 0$ is included as a diagnostic limit and is not considered when selecting the best tested β.

Effect of training-set size

In a separate experiment, we vary the number of training samples over $N_{\mathrm{tr}} \in \{400, 100, 70, 50, 30\}$ while keeping the test set fixed at $N_{\mathrm{te}} = 100$. The VAE uses its default $\beta = 1$ throughout this experiment, with no β sweep. Only reconstruction and latent balance are evaluated; generation is not considered here.

The VAE maintains consistently poor reconstruction across all tested training sizes and exhibits the least balanced latent representation, whereas RRAE and OLVAE degrade more gradually with decreasing training-set size. Table 2 summarizes the results. Because this comparison uses only the default $\beta = 1$, it should not be interpreted as representing the performance of a tuned VAE.

Table 2: Synthetic Gaussian results as the number of training samples is reduced, with $N_{\mathrm{te}} = 100$. The VAE uses $\beta = 1$. Test reconstruction MSE and latent balance $\min_j \sigma_j / \max_j \sigma_j$ are averaged over five seeds.

	Test recon. MSE ↓			Latent balance ↑		
N_{tr}	VAE (β=1)	RRAE	OLVAE	VAE (β=1)	RRAE	OLVAE
400	0.8028	0.0070	0.0105	0.663	0.820	0.897
100	0.8165	0.0750	0.0609	0.635	0.808	0.944
70	0.8124	0.0787	0.0262	0.729	0.841	0.834
50	0.8363	0.1673	0.1726	0.630	0.825	0.795
30	0.8203	0.1713	0.1340	0.730	0.910	0.897

Low-data β sensitivity

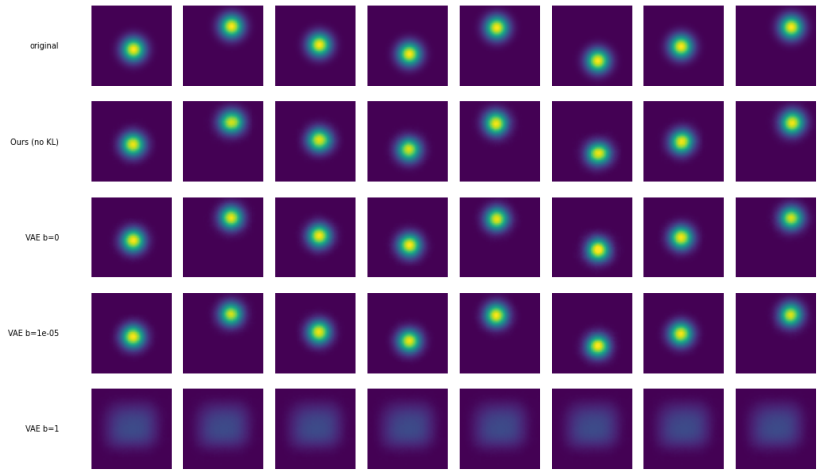
We next examine the sensitivity of the VAE to β as the amount of training data decreases, using $N_{\mathrm{tr}} \in \{100, 50, 30, 20\}$ and $N_{\mathrm{te}} = 1000$. A separate β sweep is performed for each training-set size using the same set of tested β values and five seeds. The best tested β is selected based on generation error while requiring sufficient diversity in the generated centres.

The results show that the preferred β changes with training-set size, indicating that the VAE requires re-tuning as the amount of available data changes. Interestingly, the VAE with $\beta = 0$ outperforms OLVAE at $N_{\mathrm{tr}} = 100$ and 50, whereas OLVAE performs better at $N_{\mathrm{tr}} = 30$ and 20 as the amount of training data decreases. Although the tuned VAE achieves lower generation error than OLVAE, the default $\beta = 1$ performs substantially worse across all tested sizes. In contrast, OLVAE maintains meaningful generation without requiring any β selection. Table 3 summarizes the results.

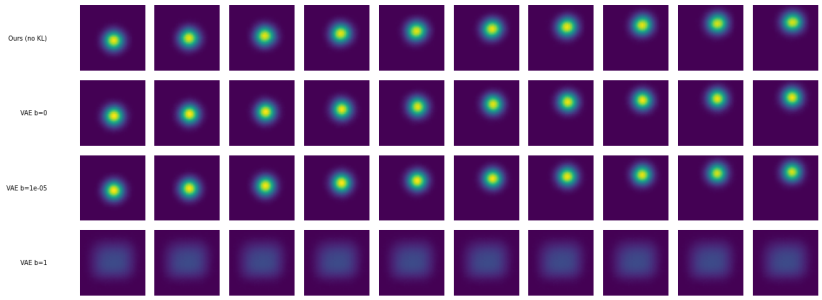
Table 3: Generation performance in the low-data regime. The VAE is tuned separately at each training-set size, with results reported over five seeds per configuration.

N_{tr}	Best VAE β	Best VAE	VAE ($\beta = 0$)	VAE ($\beta = 1$)	OLVAE
100	10^{-5}	$\mathbf{2.85}$	3.42	72.06	7.38
50	10^{-5}	$\mathbf{4.32}$	5.97	73.24	12.84
30	10^{-3}	$\mathbf{5.65}$	20.18	77.14	16.77
20	10^{-3}	$\mathbf{11.47}$	38.53	77.34	19.85

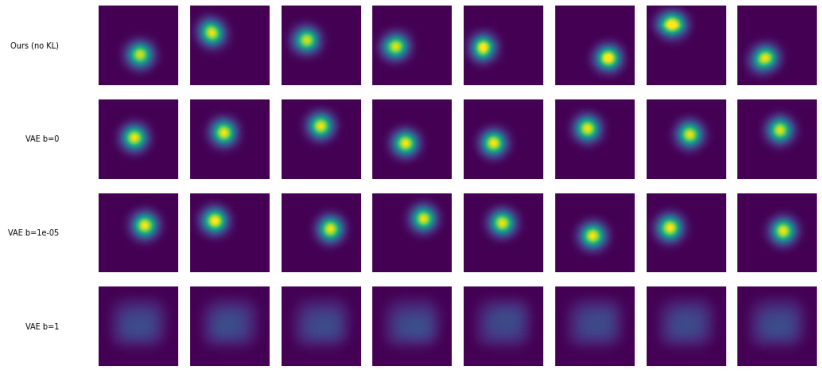
Generation error (%); lower is better.



(a) Test reconstruction



(b) Latent interpolation



(c) Random generation

Figure 2: Qualitative results on the synthetic Gaussian dataset at $N_{\mathrm{tr}} = 100$. The rows compare OLVAE with the VAE at $\beta = 0$, the best tested $\beta = 10^{-5}$, and $\beta = 1$.

3.2 Real-World Data

We evaluate OLVAE on three real-world datasets: Frey Faces, dSprites, and Shapes3D. The VAE is used as a baseline on all three datasets, while RRAE is additionally included for dSprites. Reconstruction is evaluated on held-out test data, and generation quality is assessed using FID and, for dSprites, additional distribution-based metrics.

Frey Faces

Frey Faces contains 1665 training and 300 test grayscale images, resized to 64×64. The VAE uses a ten-dimensional latent space, while OLVAE retains ten components; both are evaluated over five seeds.

VAE performance is highly sensitive to β. Generation quality varies strongly with β: very small values yield relatively high FID, intermediate values improve generation substantially, while larger values lead to severe degradation. The best tested β for generation differs from the best tested β for reconstruction, showing that the preferred value depends on the evaluation metric.

OLVAE achieves generation quality comparable to the best-generation VAE configuration while obtaining better reconstruction than that configuration, without requiring β selection. Table 4 summarizes the results.

Table 4: Frey Faces results across the tested VAE β values and OLVAE. Values are reported as mean $\pm$ sample standard deviation over five seeds.

Model	Test recon. MSE ↓	FID ↓	Inactive dims. ↓
VAE, $\beta = 0^{\dagger}$	0.0186 ± 0.0021	60.52 ± 17.79	0.0
VAE, $\beta = 10^{-6}$	0.0188 ± 0.0026	60.17 ± 13.89	0.0
VAE, $\beta = 10^{-5}$	0.0190 ± 0.0027	58.18 ± 18.18	0.0
VAE, $\beta = 10^{-4}$	$\mathbf{0.0184 \pm 0.0005}$	35.04 ± 2.44	0.0
VAE, $\beta = 10^{-3}$	0.0256 ± 0.0030	$\mathbf{27.95 \pm 1.92}$	0.0
VAE, $\beta = 10^{-2}$	0.0541 ± 0.0059	40.43 ± 2.32	0.0
VAE, $\beta = 10^{-1}$	0.3111 ± 0.0340	113.02 ± 12.00	0.8
VAE, $\beta = 1$	0.3361 ± 0.0004	121.12 ± 3.58	10.0
VAE, $\beta = 10$	0.3360 ± 0.0004	120.53 ± 1.92	10.0
VAE, $\beta = 64$	0.3361 ± 0.0004	121.12 ± 3.67	10.0
OLVAE (no β)	0.0220 ± 0.0009	28.40 ± 1.94	0.0

$^{\dagger}\beta = 0$ is included as a diagnostic limit and is not considered when selecting the best tested β.

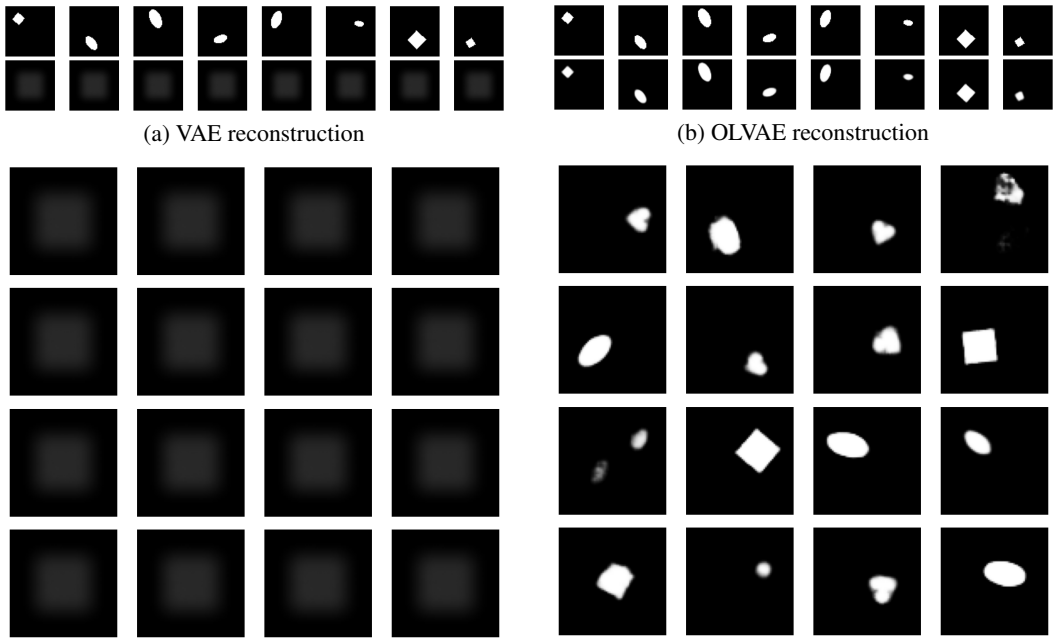
dSprites

The dSprites dataset consists of binary 64×64 images generated from five varying factors: shape, scale, orientation, and two position coordinates. We use $589{,}824$ training and $147{,}456$ test images, a five-dimensional bottleneck, and five seeds per model. Generation quality is evaluated using FID, validity rate, and Jensen–Shannon (JS) divergence. Validity measures the fraction of generated samples recognized as valid sprites, while JS divergence measures the difference between the factor distributions of real and generated samples. The VAE is evaluated at its default $\beta = 1$ in this experiment.

RRAE achieves the best reconstruction and FID on dSprites, while OLVAE achieves the highest validity rate and the lowest JS divergence. This shows that the models are ranked differently depending on the generation metric. The VAE also exhibits substantially larger variability across seeds, driven by one failed run, whereas OLVAE remains more consistent across the five seeds. The results are summarized in Table 5.

Table 5: dSprites results with a five-dimensional bottleneck. Values are reported as mean $\pm$ sample standard deviation over five seeds. The VAE uses $\beta = 1$.

Metric	VAE ($\beta = 1$)	RRAE	OLVAE
Test MSE $\downarrow$	0.00827 ± 0.01641	$\mathbf{0.00063 \pm 0.00010}$	0.00213 ± 0.00005
FID $\downarrow$	79.3 ± 73.4	$\mathbf{55.4 \pm 6.7}$	92.3 ± 1.8
Validity rate $\uparrow$	0.677 ± 0.384	0.572 ± 0.058	$\mathbf{0.953 \pm 0.015}$
Mean JS $\downarrow$	0.257 ± 0.260	0.245 ± 0.021	$\mathbf{0.139 \pm 0.011}$



(a) VAE reconstruction (b) OLVAE reconstruction

(c) VAE generation (d) OLVAE generation

Figure 3: Qualitative dSprites results for seed 1. The VAE exhibits a failed run, producing nearly identical blurry reconstructions and generations, whereas OLVAE maintains recognizable reconstructions and diverse generated samples.

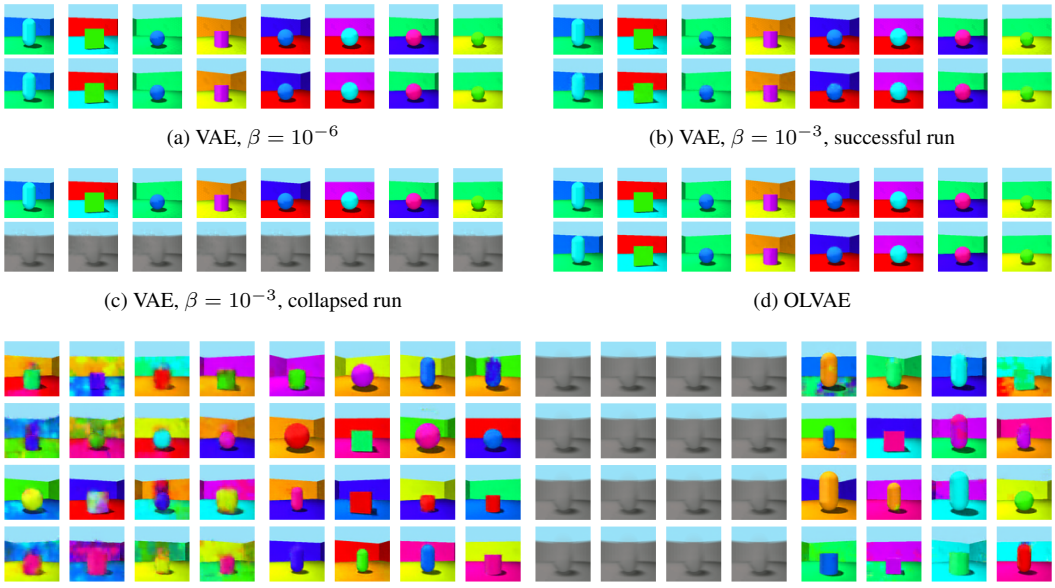
Shapes3D

We evaluate the models on the full Shapes3D dataset using $384{,}000$ training and $96{,}000$ test images. The VAE uses a 200-dimensional latent space, while OLVAE uses a 200-dimensional representation with rank 10. Both models share the same encoder–decoder architecture and training configuration and are evaluated over three random seeds. For the VAE, we evaluate multiple β values, while OLVAE does not require a KL-weight coefficient.

VAE performance on Shapes3D depends strongly on β, as shown in Table 6. A very small β yields reconstruction performance comparable to OLVAE, whereas $\beta = 1$ and $\beta = 64$ substantially degrade both reconstruction and generation. The intermediate setting $\beta = 10^{-3}$ can produce very strong samples, but its performance varies considerably across seeds. In contrast, OLVAE maintains comparable reconstruction while achieving a markedly lower and more stable FID across runs, without requiring the selection of a KL-weight coefficient.

Table 6: Shapes3D results on the full dataset. Values are reported as mean ± sample standard deviation over three seeds.

Model	Test recon. MSE ↓	FID ↓
VAE, $\beta = 10^{-6}$	$\mathbf{0.000229 \pm 0.000040}$	137.57 ± 26.54
VAE, $\beta = 10^{-3}$	0.04102 ± 0.07020	132.61 ± 161.38
VAE, $\beta = 1$	0.12279 ± 0.00137	326.32 ± 0.38
VAE, $\beta = 64$	0.12223 ± 0.00031	327.01 ± 2.50
OLVAE (no β)	0.000238 ± 0.000015	$\mathbf{47.91 \pm 3.79}$



(a) VAE, $\beta = 10^{-6}$ (b) VAE, $\beta = 10^{-3}$, successful run

(c) VAE, $\beta = 10^{-3}$, collapsed run (d) OLVAE

(e) VAE, $\beta = 10^{-6}$ (f) VAE, $\beta = 10^{-3}$ (successful) (g) VAE, $\beta = 10^{-3}$ (collapsed) (h) OLVAE

Figure 4: Qualitative Shapes3D results. Panels (a–d) show test reconstructions, while panels (e–h) show random generations for VAE and OLVAE.

4 Ablation Study

The results presented in the previous sections show that OLVAE achieves a favorable trade-off between reconstruction and generation quality. In this section, we conduct four ablations to examine: (i) the individual contributions of the truncated SVD and the proposed G_B transformation, (ii) the specific construction of G_B, (iii) the noise scale σ, and (iv) the mini-batch size over which the SVD and coefficient mixing are performed. Unless stated otherwise, all models use the same encoder–decoder backbone, optimizer, and training budget, and results are reported as mean ± standard deviation over five random seeds.

4.1 Component Ablation: SVD, G_B, and Their Combination

We first isolate the contributions of the two main mechanisms in Eqs. (4)–(6). We compare: *SVD only*, for which the truncated representation $\bar{Y}$ is passed directly to the decoder with $G_B = I_B$; G_B *only*, for which the truncated-SVD step is removed and G_B is right-multiplied directly onto the encoder output Y; and *SVD+*G_B, corresponding to the complete OLVAE model. All variants are trained using only the reconstruction objective in Eq. (10).

Removing the SVD also removes the explicit rank-k constraint. We therefore evaluate the G_B-only variant under two latent widths: $d_z = k$, which provides a matched bottleneck, and $d_z = 200$, which retains the wide latent width used before truncation in the SVD-based variants. The same network

backbone, optimizer, and training budget are used, with the latent width and rank-reduction/mixing mechanism changed only where required.

For SVD-based variants, random generation follows Eq. (13) using the frozen low-rank basis. Since the G_B-only variant has no SVD basis, its latent codes are sampled directly as $z \sim \mathcal{N}(0, I_{d_z})$ and passed to the decoder. For the synthetic Gaussian dataset, we use $k = 2$, $N_{\mathrm{tr}} = 400$, and $N_{\mathrm{te}} = 100$, and evaluate generation using the Gaussian-fit relative error. For Frey Faces, $k = 10$ and generation is evaluated using FID.

Table 7: Component ablation of the truncated SVD and G_B transformation. Lower values are better. Results are mean $\pm$ standard deviation over five seeds.

Dataset	Variant	d_z	Test recon. $\downarrow$	Generation metric $\downarrow$
Gaussian	SVD only	200	0.0103 ± 0.0093	$11.31 \pm 7.64\%$
Gaussian	G_B only	$k = 2$	0.0083 ± 0.0015	$3.92 \pm 0.82\%$
Gaussian	G_B only	200	0.0031 ± 0.0006	$90.21 \pm 6.76\%$
Gaussian	SVD+G_B (OLVAE)	200	0.0169 ± 0.0068	$4.68 \pm 0.60\%$
Frey Faces	SVD only	200	0.0179 ± 0.0008	93.0 ± 13.1
Frey Faces	G_B only	$k = 10$	0.0202 ± 0.0010	30.5 ± 3.4
Frey Faces	G_B only	200	0.0163 ± 0.0004	53.2 ± 3.3
Frey Faces	SVD+G_B (OLVAE)	200	0.0217 ± 0.0010	30.9 ± 4.1

SVD only gives competitive reconstruction errors but substantially worse generation quality on both datasets. At the matched bottleneck ($d_z = k$), G_B only already provides a large improvement in generation over SVD only, showing that the proposed stochastic transformation contributes substantially to the generative behavior. OLVAE achieves similarly strong generation, while giving the lowest generation-error variability on the Gaussian dataset.

Increasing the G_B-only latent width from $d_z = k$ to $d_z = 200$ further improves reconstruction but sharply degrades generation. In this setting, generation is performed directly from an unconstrained 200-dimensional isotropic prior, whereas the SVD-based variants generate through the learned frozen low-rank basis. These results support complementary roles for the two mechanisms: the truncated SVD provides an explicit rank-k latent structure, while G_B introduces structured stochastic perturbations of the retained coefficients.

4.2 G_B-Design Ablation

We next investigate whether the specific construction of G_B is important for generative performance. Four variants are compared. The proposed *signed-noise* construction uses Eq. (5). The *positive-noise* variant replaces Ξ by $|\Xi|$. The *softmax-mixing* variant is constructed from $I_B + \frac{\sigma}{\sqrt{B}}\Xi$ followed by column-wise softmax normalization. Finally, the *Gaussian-kernel* variant constructs G_B deterministically from pairwise distances between the retained coefficient vectors within the current mini-batch.

Thus, the first three variants are derived from the same Gaussian random matrix Ξ, whereas the Gaussian-kernel construction is data-dependent and contains no random Ξ. All variants are applied after the truncated SVD and are trained under the same experimental settings. Since this ablation examines the generative behavior of the different G_B constructions, we report Gaussian-fit relative error for the synthetic dataset and FID for Frey Faces.

The signed construction achieves the lowest generation error on the Gaussian dataset and the lowest FID on Frey Faces. Restricting the perturbation to positive values degrades generation, while the softmax and Gaussian-kernel variants perform progressively worse. Reconstruction error does not follow the same ordering on Frey Faces: the Gaussian-kernel variant obtains the lowest reconstruction error among the four variants despite having the worst FID. This further illustrates that reconstruction quality alone does not imply good generative behavior. We therefore use the signed, identity-centred formulation of G_B in the remaining experiments.

Table 8: G_B-design ablation. Lower values are better. Results are mean $\pm$ standard deviation over five seeds.

G_B design	Gaussian-fit error (%) $\downarrow$	Frey FID $\downarrow$
Signed noise (ours)	$\mathbf{7.33 \pm 2.27}$	$\mathbf{27.63 \pm 1.34}$
Positive noise	12.64 ± 4.35	43.91 ± 3.58
Softmax mixing	17.41 ± 12.38	67.06 ± 9.64
Gaussian kernel	32.06 ± 42.16	95.09 ± 5.69

Table 9: Batch-size ablation on the synthetic Gaussian dataset ($N_{\text{tr}} = 400$). Lower values are better.

B	Test recon. $\downarrow$	Gaussian-fit error (%) $\downarrow$
16	0.0140 ± 0.0049	4.57 ± 1.38
32	0.0163 ± 0.0113	4.57 ± 1.48
64	0.0194 ± 0.0107	5.31 ± 1.08
128	0.0127 ± 0.0027	6.12 ± 1.50

4.3 Noise-Scale (σ) Ablation

We study how the noise scale σ in G_B affects reconstruction and generation quality, to justify the choice of $\sigma = 0.1$ used throughout the paper. We sweep $\sigma \in \{0.05, 0.1, 0.2, 0.5, 1.0\}$ at the signed-noise G_B design, on the same $N_{tr} = 400$ Gaussian setting ($k = 2$) and Frey Faces setting ($k = 10$) used in the preceding ablations, over five seeds, and report mean $\pm$ std for test reconstruction error and for the corresponding generation metric (Gaussian-fit relative error on the synthetic dataset, FID on Frey Faces).

Reconstruction error increases with σ on both datasets, roughly monotonically, from 0.0157 ± 0.0153 at $\sigma = 0.05$ to 0.3907 ± 0.0368 at $\sigma = 1.0$ on the Gaussian dataset, and from 0.0189 ± 0.0007 to 0.1203 ± 0.0121 on Frey Faces over the same range. Generation quality, in contrast, is not monotonic in σ. On Frey Faces, FID is worst at $\sigma = 0$ (SVD-only reference, 93.0 ± 13.1, from Section **??**), falls sharply to its lowest value at $\sigma = 0.1$ (27.69 ± 1.26), and rises again for larger σ, reaching 56.96 ± 3.83 at $\sigma = 1.0$. The Gaussian dataset shows the same pattern once its own $\sigma = 0$ reference is included ($11.31 \pm 7.64\%$): generation error falls to its lowest values at $\sigma = 0.05$–0.1 (3.95–4.71%) and then increases steadily for $\sigma \geq 0.2$, up to $48.13 \pm 2.38\%$ at $\sigma = 1.0$. This indicates that some stochasticity in G_B is needed for good generation — consistent with the component ablation, where the SVD-only variant had the worst generation quality on both datasets — but that generation quality degrades again once σ grows large enough to substantially harm reconstruction. Within the range tested, $\sigma = 0.1$ lies at, or immediately next to, the best generation quality on both datasets, which is why it is adopted as the default noise scale for the rest of the paper.

4.4 Batch-Size Ablation

Since both the truncated SVD and G_B operate on the current mini-batch, we investigate sensitivity to the batch size $B \in \{16, 32, 64, 128\}$ while keeping the retained rank, σ, and the remaining training settings fixed.

Reconstruction error shows no consistent dependence on batch size on either dataset. Generation quality, however, degrades as the batch size increases. On Frey Faces, FID increases monotonically from 22.6 ± 0.7 at $B = 16$ to 56.0 ± 8.0 at $B = 128$, corresponding to an approximately $2.5\times$ increase. The Gaussian-fit error shows a weaker but similar trend, remaining at 4.57% for $B = 16$ and $B = 32$ before increasing to 5.31% and 6.12% for $B = 64$ and $B = 128$, respectively.

These results indicate a batch-size dependence of the current batch-wise SVD and coefficient-mixing mechanism, with generation being substantially more sensitive than reconstruction. Smaller batches, particularly $B = 16$–32, therefore provide the most favorable generative performance among the tested settings. Sensitivity to large batch sizes is considered a limitation of the current formulation.

4.5 Retained-Rank Ablation

We examine the sensitivity of OLVAE to the retained rank k on Frey Faces and Shapes3D, with the wide latent fixed at $d_z = 200$ and all other settings (signed G_B, $\sigma = 0.1$, optimizer, training budget)

Table 10: Batch-size ablation on Frey Faces. Lower values are better.

B	Test recon. ↓	FID ↓
16	0.0230 ± 0.0003	22.6 ± 0.7
32	0.0214 ± 0.0008	27.9 ± 0.8
64	0.0208 ± 0.0005	39.1 ± 2.4
128	0.0213 ± 0.0009	56.0 ± 8.0

Table 11: Retained-rank ablation on Frey Faces and Shapes3D. The wide latent dimension is fixed at $d_z = 200$; lower is better for all columns. Frey Faces is evaluated by reconstruction only; Shapes3D additionally by FID.

Retained rank k	Frey recon ↓	Shapes3D recon ↓	Shapes3D FID ↓
2	0.1142	0.02273	167.4
5	0.0482	0.00210	57.3
6	–	0.00087	50.1
10	0.0214	0.00025	47.8
20	0.0176	0.00018	113.0
30	–	0.00020	183.9

unchanged. Each rank sweep is trained at a single fixed nominal seed, so k is the primary factor varied; we report a single run per rank.

On Frey Faces, reconstruction improves consistently with k, with most of the gain realized by $k = 10$; increasing k to 20 gives comparatively little further improvement (Table 11). We therefore fix $k = 10$ for Frey Faces as a compact rank that retains most of the benefit of a wider bottleneck.

On Shapes3D, reconstruction again improves consistently with k up to $k = 20$, after which it changes little. FID, in contrast, is clearly non-monotonic: it falls sharply up to $k = 10$, where it reaches its minimum, then rises again at $k = 20$ and further at $k = 30$, even though reconstruction continues to improve or stay flat over that range. The best generative performance is therefore obtained at an intermediate rank rather than the rank that minimizes reconstruction error — increasing k beyond what reconstruction requires does not improve, and can substantially degrade, generation quality.

5 Discussion

The ablations show that OLVAE relies on two complementary mechanisms: truncated SVD provides low-rank latent structure, while the multiplicative Gaussian transformation introduces stochasticity needed for generation.The component ablation indicates that the two mechanisms play complementary roles in the full model. The β-sweeps show that the VAE's best β changes with the dataset, the evaluation metric, and the training-set size, which is exactly the tuning burden that OLVAE removes. In the cases where a well-tuned VAE still edges out OLVAE for example, at the higher training-set sizes in the Gaussian experiments, and dSprites reconstruction where RRAE leads), OLVAE trades a small amount of peak performance for robustness and the elimination of β.

6 Limitations and Future Work

The main limitation is a dependence on mini-batch size. Because the SVD and coefficient mixing are computed on the current batch, generation quality degrades for large batches (FID rises roughly $2.5\times$ from B=16 to B=128 on Frey Faces), so small batches are currently preferred. Future work includes reducing this batch-size dependence, for example with a running or global basis, extending OLVAE to higher-resolution and more varied datasets, and a fuller theoretical characterization of the conditional Gaussian latent structure.

7 Conclusion

We introduced OLVAE, a generative autoencoder that replaces the VAE's KL-divergence regularization with a truncated SVD followed by a signed, identity-centred multiplicative Gaussian

transformation of the retained coefficients. This yields a stochastic latent space that supports reconstruction, interpolation, and random generation while training with a single reconstruction loss, without a KL term or a balancing coefficient such as β. Across four datasets, OLVAE is competitive with a VAE, and better in several settings, while avoiding the per-dataset, per-metric β search. Its advantage is clearest in the low-data regime and on Shapes3D, where the VAE is highly β-sensitive and unstable across seeds. The main remaining limitation is sensitivity to mini-batch size. Overall, OLVAE shows that low-rank structure combined with structured stochastic noise is a practical and KL-free alternative to KL-regularized generative autoencoders.

References

David M. Blei, Alp Kucukelbir, and Jon D. McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017. doi: 10.1080/01621459.2017.1285773.

Samuel R. Bowman, Luke Vilnis, Oriol Vinyals, Andrew M. Dai, Rafal Jozefowicz, and Samy Bengio. Generating sentences from a continuous space. In *SIGNLL Conference on Computational Natural Language Learning (CoNLL)*, pages 10–21, 2016. doi: 10.18653/v1/K16-1002.

Chris Burgess and Hyunjik Kim. 3D Shapes dataset. `https://github.com/google-deepmind/3d-shapes`, 2018.

Steffen Czolbe, Oswin Krause, Ingemar Cox, and Christian Igel. A loss function for generative neural networks based on Watson's perceptual model. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, 2020. URL `https://proceedings.neurips.cc/paper/2020/hash/165a59f7cf3b5c4396ba65953d679f17-Abstract.html`.

Brendan J. Frey. Frey face dataset. `https://cs.nyu.edu/home/people/in_memoriam/roweis/data.html`, n.d. Distributed by Sam Roweis; accessed 29 July 2026.

Hao Fu, Chunyuan Li, Xiaodong Liu, Jianfeng Gao, Asli Celikyilmaz, and Lawrence Carin. Cyclical annealing schedule: A simple approach to mitigating KL vanishing. In *Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*, pages 240–250, 2019. doi: 10.18653/v1/N19-1021.

Rafael Gómez-Bombarelli, Jennifer N. Wei, David Duvenaud, José Miguel Hernández-Lobato, Benjamín Sánchez-Lengeling, Dennis Sheberla, Jorge Aguilera-Iparraguirre, Timothy D. Hirzel, Ryan P. Adams, and Alán Aspuru-Guzik. Automatic chemical design using a data-driven continuous representation of molecules. *ACS Central Science*, 4(2):268–276, 2018. doi: 10.1021/acscentsci.7b00572.

Junxian He, Daniel Spokoyny, Graham Neubig, and Taylor Berg-Kirkpatrick. Lagging inference networks and posterior collapse in variational autoencoders. In *International Conference on Learning Representations (ICLR)*, 2019. URL `https://openreview.net/forum?id=rylDfnCqF7`.

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 (ICLR)*, 2017. URL `https://openreview.net/forum?id=Sy2fzU9gl`.

Diederik P. Kingma and Max Welling. An introduction to variational autoencoders. *Foundations and Trends in Machine Learning*.

Diederik P. Kingma and Max Welling. Auto-encoding variational Bayes. In *International Conference on Learning Representations (ICLR)*, 2014. URL `https://arxiv.org/abs/1312.6114`.

Diederik P. Kingma, Tim Salimans, Rafal Jozefowicz, Xi Chen, Ilya Sutskever, and Max Welling. Improved variational inference with inverse autoregressive flow. In *Advances in Neural Information Processing Systems*, volume 29, pages 4743–4751, 2016. URL `https://proceedings.neurips.cc/paper/2016/hash/ddeebdeefdb7e7e7a697e1c3e3d8ef54-Abstract.html`.

Anders Boesen Lindbo Larsen, Søren Kaae Sønderby, Hugo Larochelle, and Ole Winther. Autoencoding beyond pixels using a learned similarity metric. In *International Conference on Machine Learning (ICML)*, pages 1558–1566, 2016. URL `https://proceedings.mlr.press/v48/larsen16.html`.

Romain Lopez, Jeffrey Regier, Michael B. Cole, Michael I. Jordan, and Nir Yosef. Deep generative modeling for single-cell transcriptomics. *Nature Methods*, 15(12):1053–1058, 2018. doi: 10.1038/s41592-018-0229-2.

James Lucas, George Tucker, Roger B. Grosse, and Mohammad Norouzi. Don't blame the ELBO! a linear VAE perspective on posterior collapse. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/7e3315fe390974fcf25e44a9445bd821-Abstract.html.

Loic Matthey, Irina Higgins, Demis Hassabis, and Alexander Lerchner. dSprites: Disentanglement testing sprites dataset. https://github.com/google-deepmind/dsprites-dataset, 2017.

Volodymyr Melnykov and Ranjan Maitra. Finite mixture models and model-based clustering. *Statistics Surveys*, 4:80–116, 2010. doi: 10.1214/09-SS053.

Jad Mounayer, Sebastian Rodriguez, Chady Ghnatios, Charbel Farhat, and Francisco Chinesta. Rank reduction autoencoders. *arXiv preprint arXiv:2405.13980*, 2024. URL https://arxiv.org/abs/2405.13980.

Jad Mounayer, Alicia Tierz, Jerome Tomezyk, Chady Ghnatios, and Francisco Chinesta. Variational rank reduction autoencoders. *arXiv preprint arXiv:2505.09458*, 2025. URL https://arxiv.org/abs/2505.09458.

Lawrence R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989. doi: 10.1109/5.18626.

Ali Razavi, Aäron van den Oord, Ben Poole, and Oriol Vinyals. Preventing posterior collapse with delta-VAEs. In *International Conference on Learning Representations (ICLR)*, 2019. URL https://arxiv.org/abs/1901.03416.

Ilya Tolstikhin, Olivier Bousquet, Sylvain Gelly, and Bernhard Schölkopf. Wasserstein auto-encoders. In *International Conference on Learning Representations (ICLR)*, 2018. URL https://openreview.net/forum?id=HkL7n1-0b.

Jakub M. Tomczak and Max Welling. VAE with a VampPrior. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 84 of *Proceedings of Machine Learning Research*, pages 1214–1223. PMLR, 2018. URL https://proceedings.mlr.press/v84/tomczak18a.html.

Sebastian J. Wetzel. Unsupervised learning of phase transitions: From principal component analysis to variational autoencoders. *Physical Review E*, 96(2):022140, 2017. doi: 10.1103/PhysRevE.96.022140.

Shengjia Zhao, Jiaming Song, and Stefano Ermon. InfoVAE: Balancing learning and inference in variational autoencoders. In *AAAI Conference on Artificial Intelligence*, pages 5885–5892, 2019. doi: 10.1609/aaai.v33i01.33015885.