Supplementary Material
*dp*VAEs: Fixing Sample Generation for Regularized VAEs

Riddhish Bhalodia, Iain Lee, Shireen Elhabian

Scientific Computing and Imaging Institute, University of Utah

Abstract. This supplementary material includes the derivation of *dp*VAE training objective, the ELBO definitions of state-of-the-art VAE regularizers with the decoupled prior, and experimental details (architecture, hyperparameters, and train/test splits) for MNIST, SVHN and CelebA experiments. It also includes the quantitative results that showcase that regularization benefit of the disentanglement inducing regularizers does not diminish significantly when the decouple priors are added. Finally, it showcases the generation results with different regularization with and without the decoupled prior on MNIST and Celeb-A, along with the latent traversal results on Celeb-A.

1 Derivation of KL $[q_\phi(\mathbf{z}|\mathbf{x})\|p(\mathbf{z})]$ with Decoupled Prior

Consider a bijective mapping between $\mathcal{Z}$ and $\mathcal{Z}_0$ defined by a function $g_\eta(\mathbf{z}) = \mathbf{z}_0$. The *change of variable formula* for mapping probability distribution on $\mathbf{z}$ to $\mathbf{z}_0$ is given as follows:

$$p(\mathbf{z}) = p(\mathbf{z}_0)\left|\frac{\partial \mathbf{z}_0}{\partial \mathbf{z}}\right| = p(g_\eta(\mathbf{z}))\left|\frac{\partial g_\eta(\mathbf{z})}{\partial \mathbf{z}}\right| \tag{1}$$

The $g-$bijection is parameterized by K *affine coupling layers*, each layer is a bijection block $\mathbf{z}_{k-1} = g_\eta^{(k)}(\mathbf{z}_k)$ of the form,

$$\begin{aligned} g_\eta^{(k)}(\mathbf{z}_k) = \mathbf{b}_k \odot \mathbf{z}_k \\ + (1 - \mathbf{b}_k) \odot [\mathbf{z}_k \odot \exp\left(s_k(\mathbf{b}_k \odot \mathbf{z}_k)\right) \\ + t_k\left(\mathbf{b}_k \odot \mathbf{z}_k\right)], \end{aligned} \tag{2}$$

where $\mathbf{z} = \mathbf{z}_K$, $\odot$ is the Hadamard (*i.e.,* element-wise) product, $\mathbf{b}_k \in \{0,1\}^L$ is a binary mask used for partitioning the $k-$th block input, and $\eta = \{s_1, ..., s_K, t_1, ..., t_K\}$ are the deep networks parameters of the scaling s_k and translation t_k functions of the K blocks. Stacking these affine coupling layers constitutes the functional mapping between the representation and the generation spaces. The $g-$bijection is thus defined as,

$$g_\eta(\mathbf{z}) = \mathbf{z}_0 = g_\eta^{(1)} \circ \cdots \circ g_\eta^{(K-1)} \circ g_\eta^{(K)}(\mathbf{z}). \tag{3}$$

The KL divergence is given as,

$$\begin{aligned} \mathrm{KL} &:= \mathrm{KL}\left[q_\phi(\mathbf{z}|\mathbf{x})\|p(\mathbf{z})\right] \\ &= \int q_\phi(\mathbf{z}|\mathbf{x})\left[\log(q_\phi(\mathbf{z}|\mathbf{x})) - \log(p(\mathbf{z}))\right] d\mathbf{z}. \end{aligned} \tag{4}$$

Using the change of variable formula in (1), we have the following.

$$\begin{aligned}\mathrm{KL} &= \int q_\phi(\mathbf{z}|\mathbf{x}) \Bigg[\log(q_\phi(\mathbf{z}|\mathbf{x})) - \log(p(g_\eta(\mathbf{z}))) \\ &\qquad - \log\left(\left|\frac{\partial g_\eta(\mathbf{z})}{\partial \mathbf{z}}\right|\right)\Bigg] d\mathbf{z} \qquad (5)\\ &= \mathrm{KL}\left[q_\phi(\mathbf{z}|\mathbf{x}) \| p(g_\eta(\mathbf{z}))\right] \\ &\qquad - \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log\left(\left|\frac{\partial g_\eta(\mathbf{z})}{\partial \mathbf{z}}\right|\right)\right]. \qquad (6)\end{aligned}$$

The first term in (6) can be derived as follows,

$$\begin{aligned}T_1 &= \mathrm{KL}\left[q_\phi(\mathbf{z}|\mathbf{x}) \| p(g_\eta(\mathbf{z}))\right] \qquad (7)\\ &= \int q_\phi(\mathbf{z}|\mathbf{x}) \log\left(\frac{q_\phi(\mathbf{z}|\mathbf{x})}{p(g_\eta(\mathbf{z}))}\right) d\mathbf{z} \qquad (8)\\ &= -\mathbb{H}(q_\phi(\mathbf{z}|\mathbf{x})) - \int q_\phi(\mathbf{z}|\mathbf{x}) \log\left(p(g_\eta(\mathbf{z}))\, d\mathbf{z}\right), \qquad (9)\end{aligned}$$

where $\mathbb{H}(q_\phi(\mathbf{z}|\mathbf{x}))$ is the differential entropy of the variational posterior distribution. This approximate posterior is a multivariate Gaussian with a diagonal covariance matrix, *i.e.,* $q_\phi(\mathbf{z}|\mathbf{x}) \sim \mathcal{N}(\mathbf{z}; \boldsymbol{\mu}_\mathbf{z}(\mathbf{x}), \boldsymbol{\Sigma}_\mathbf{z}(\mathbf{x}))$, where $\boldsymbol{\Sigma}_\mathbf{z}(\mathbf{x}) = \mathrm{diag}(\boldsymbol{\sigma}_\mathbf{z}(\mathbf{x}))$, and $\boldsymbol{\sigma}_\mathbf{z}(\mathbf{x}) \in \mathbb{R}_+^L$. This results in a closed-form expression for the entropy of the approximate posterior [1], given as:

$$\mathbb{H}(q_\phi(\mathbf{z}|\mathbf{x})) = \frac{L}{2} + \frac{L}{2}\log(2\pi) + \frac{\log|\boldsymbol{\Sigma}_\mathbf{z}(\mathbf{x})|}{2}. \qquad (10)$$

The probability of the base latent space (*i.e.,* the generation space) in the decoupled prior is assumed to be a standard normal distribution, *i.e.,* $p(g_\eta(\mathbf{z})) = \mathcal{N}(g_\eta(\mathbf{z}); 0, \mathbb{I}_L)$. Together with (10) and ignoring constants terms, the first term in (6) can be simplified into,

$$\begin{aligned}T_1 &= -\frac{\log|\boldsymbol{\Sigma}_\mathbf{z}(\mathbf{x})|}{2} + \int q_\phi(\mathbf{z}|\mathbf{x}) \frac{L}{2} \log\left(\frac{1}{2\pi}\right) d\mathbf{z} \\ &\quad + \frac{1}{2}\int q_\phi(\mathbf{z}|\mathbf{x}) g_\eta(\mathbf{z})^T g_\eta(\mathbf{z}) d\mathbf{z} \qquad (11)\\ &= \frac{1}{2}\Bigg[-\log|\boldsymbol{\Sigma}_\mathbf{z}(\mathbf{x})| + const \\ &\quad + \mathbb{E}_{q(\mathbf{z}|\mathbf{x})}\left[g_\eta(\mathbf{z})^T g_\eta(\mathbf{z})\right]\Bigg] \qquad (12)\end{aligned}$$

The second term in (6) can be derived as follows,

$$\begin{aligned}T_2 &= \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log\left(\left|\frac{\partial g_\eta(\mathbf{z})}{\partial \mathbf{z}}\right|\right)\right] \\ &= \int q_\phi(\mathbf{z}|\mathbf{x}) \log\left(\left|\frac{\partial g_\eta(\mathbf{z})}{\partial \mathbf{z}}\right|\right) d\mathbf{z}. \qquad (13)\end{aligned}$$

Applying the chain rule to $g_\eta(\mathbf{z})$, which is a composition of functions as defined in (3), and combining this with the multiplicative property of determinants, we

have the following,

$$\left|\frac{\partial g_\eta(\mathbf{z})}{\partial \mathbf{z}}\right| = \prod_{k=1}^{K} \left|\frac{\partial g_\eta^{(k)}(\mathbf{z}_k)}{\partial \mathbf{z}_k}\right|. \tag{14}$$

Hence, we have the second term in (6) can be expressed as follows:

$$T_2 = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\sum_{k=1}^{K} \log\left(\left|\frac{\partial g_\eta^{(k)}(\mathbf{z}_k)}{\partial \mathbf{z}_k}\right|\right)\right]. \tag{15}$$

Similar to [2], deriving the Jacobian of individual affine coupling layers yields to an upper triangular matrix. Hence, the determinant is simply the product of the diagonal values, resulting in,

$$\log\left(\left|\frac{\partial g_\eta^{(k)}(\mathbf{z}_k)}{\partial \mathbf{z}_k}\right|\right) = \sum_{l=1}^{L} b_k^l s_k(b_k^l z_k^l). \tag{16}$$

Using this simplification, we finally have an expression for the second term expressed as,

$$T_2 = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\sum_{k=1}^{K}\sum_{l=1}^{L} b_k^l s_k(b_k^l z_k^l)\right]. \tag{17}$$

Substituting (17) and (12) in (6), the KL divergence term of the decoupled prior can be written as,

$$\begin{aligned}\mathrm{KL} = \frac{1}{2}\Big[&- \log|\boldsymbol{\Sigma}_\mathbf{z}(\mathbf{x})| + \mathbb{E}_{q(\mathbf{z}|\mathbf{x})}\left[g_\eta(\mathbf{z})^T g_\eta(\mathbf{z})\right]\Big] \\ &- \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\sum_{k=1}^{K}\sum_{l=1}^{L} b_k^l s_k(b_k^l z_k^l)\right]\end{aligned} \tag{18}$$

2 ELBOs for Different Regularizers

In this section, we define the ELBO (*i.e.,* training objective) for individual regularizers considered (see section 4.3 in the main manuscript) and how they differ under the application of decoupled prior. Here, we only serve to provide more mathematical clarity of these modifications.

β-*dp*VAE: The ELBO for the β-VAE [3] can be expressed as follows:

$$\mathcal{L}(\theta,\phi) = \mathbb{E}_{p(\mathbf{x})}\left[\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log p_\theta(\mathbf{x}|\mathbf{z})\right] - \beta\,\mathrm{KL}\left[q_\phi(\mathbf{z}|\mathbf{x})\|p(\mathbf{z})\right]\right] \tag{19}$$

By substituting the KL divergence under the decoupled prior, the ELBO for β-*dp*VAE can defined as follows:

$$\begin{aligned}\mathcal{L}(\theta,\phi,\eta) = \mathbb{E}_{p(\mathbf{x})}\Bigg[&\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log p_\theta(\mathbf{x}|\mathbf{z})\right] \\ &- \frac{\beta}{2}\Big[- \log|\boldsymbol{\Sigma}_\mathbf{z}(\mathbf{x})| + \mathbb{E}_{q(\mathbf{z}|\mathbf{x})}\left[g_\eta(\mathbf{z})^T g_\eta(\mathbf{z})\right]\Big] \\ &- \beta\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\sum_{k=1}^{K}\sum_{l=1}^{L} b_k^l s_k(b_k^l z_k^l)\right]\Bigg]\end{aligned} \tag{20}$$

The alternate formulation, β-VAE-B [4], has a similar ELBO function with some additional parameters applied to the KL divergence term.

Factor-*dp*VAE: The ELBO for FactorVAE [5] is given as follows:

$$\begin{aligned}\mathcal{L}(\theta,\phi) = \mathbb{E}_{p(\mathbf{x})}\left[\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log p_\theta(\mathbf{x}|\mathbf{z})\right] - \mathrm{KL}\left[q_\phi(\mathbf{z}|\mathbf{x})\|p(\mathbf{z})\right]\right] \\ - \gamma\,\mathrm{KL}\left[q_\phi(\mathbf{z})\|q_\phi(\bar{\mathbf{z}})\right] \end{aligned} \tag{21}$$

To modifying this ELBO with the decoupled prior, the KL divergence term between the posterior and prior that contains $p(\mathbf{z})$ is the only term that needs to be reformulated. This results in:

$$\begin{aligned}\mathcal{L}(\theta,\phi,\eta) = \mathbb{E}_{p(\mathbf{x})}\Bigg[\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log p_\theta(\mathbf{x}|\mathbf{z})\right] \\ - \frac{1}{2}\left[-\log|\boldsymbol{\Sigma}_{\mathbf{z}}(\mathbf{x})| + \mathbb{E}_{q(\mathbf{z}|\mathbf{x})}\left[g_\eta(\mathbf{z})^T g_\eta(\mathbf{z})\right]\right] \\ - \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\sum_{k=1}^{K}\sum_{l=1}^{L} b_k^l s_k(b_k^l z_k^l)\right]\Bigg] \end{aligned} \tag{22}$$

$$- \gamma\,\mathrm{KL}\left[q_\phi(\mathbf{z})\|q_\phi\bar{(}\mathbf{z})\right] \tag{23}$$

β-TC-*dp*VAE: Following similar notation as provided in [6], the ELBO for β-TCVAE [6] is given as follows:

$$\begin{aligned}\mathcal{L}(\theta,\phi) = \mathbb{E}_{p(n)}\Bigg[\mathbb{E}_{q_\phi(\mathbf{z}|n)}\left[\log p_\theta(n|\mathbf{z})\right]] \\ - \alpha I_{q_\phi}(\mathbf{z};n) - \beta\,\mathrm{KL}\left[q_\phi(\mathbf{z})\|\prod_j q_\phi(z_j)\right]\Bigg] \\ - \gamma\sum_j \mathrm{KL}\left[q_\phi(z_j)\|p(z_j)\right] \end{aligned} \tag{24}$$

Due to the factorized representation of the prior, the last KL divergence is computed via sampling. Hence, the modification for the decoupled prior becomes trivial. We simply sample from $\mathbf{z}_0$ space (which is assumed to be factorized) and pass it through $g_\eta^{-1}(z_{0j})$ to obtain a sample in $\mathbf{z}$ space, the ELBO can thus be modified as follows:

$$\begin{aligned}\mathcal{L}(\theta,\phi,\eta) = \mathbb{E}_{p(n)}\left[\mathbb{E}_{q_\phi(\mathbf{z}|n)}\left[\log p_\theta(n|\mathbf{z})\right]] - \alpha I_{q_\phi}(\mathbf{z};n)\right] \\ - \beta\,\mathrm{KL}\left[q_\phi(\mathbf{z})\|\prod_j q_\phi(z_j)\right] - \gamma\sum_j \mathrm{KL}\left[q_\phi(z_j)\|p(g_\eta^{-1}(z_{0j}))\right] \end{aligned} \tag{25}$$

Info-*dp*VAE: For InfoVAE [7], the ELBO (using the MMD divergence) is given as follows:

$$\begin{aligned}\mathcal{L}(\theta,\phi) = \mathbb{E}_{p(\mathbf{x})}\left[\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log p_\theta(\mathbf{x}|\mathbf{z})\right] - (1-\alpha)\,\mathrm{KL}\left[q_\phi(\mathbf{z}|\mathbf{x})\|p(\mathbf{z})\right]\right] \\ - (\alpha+\lambda-1)D_{MMD}(q_\phi(\mathbf{z})||p(\mathbf{z})) \end{aligned} \tag{26}$$

The modification for the decoupled prior takes place in two terms; the MMD divergence term, which is computed via sampling from the aggregate posterior, and the prior in $\mathbf{z}$ space, which acts on the representation space in the decoupled prior. Therefore, the modification is very similar to the one applied in β-TCVAE. Additionally, the KL divergence term will be modified normally. The final ELBO is thus as follows:

$$\begin{aligned}\mathcal{L}(\theta,\phi,\eta) = \mathbb{E}_{p(\mathbf{x})}\Bigg[&\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log p_\theta(\mathbf{x}|\mathbf{z})\right] - \frac{1-\alpha}{2}\Big[-\log|\boldsymbol{\Sigma}_{\mathbf{z}}(\mathbf{x})| \\ &+ \mathbb{E}_{q(\mathbf{z}|\mathbf{x})}\left[g_\eta(\mathbf{z})^T g_\eta(\mathbf{z})\right]\Big] \\ &- (1-\alpha)\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\sum_{k=1}^{K}\sum_{l=1}^{L} b_k^l s_k(b_k^l z_k^l)\right]\Bigg] \\ &- D_{MMD}\left(q_\phi(\mathbf{z})||p(g_\eta^{-1}(\mathbf{z}_0))\right)\end{aligned} \tag{27}$$

3 Architectures and Hyperparameters

In this section, we give more details for the architecture and hyperparameters used and the data handling for the four different datasets used in the paper, namely two moons toy data, MNIST, SVHN, and CelebA. These are also described for different regularizers used on each of these datasets.

Figure 1 illustrates the VAE architecture for all the datasets and the regularizers reported in the experiments section of the paper. For MNIST and SVHN, we use ReLU as the non-linear activation function, and for CelebA we use leaky ReLU [8]. For the two moons data, the VAE architecture consists of two fully connected layers of size 100 and 50 (from input to latent space) in the encoder. The decoder is a mirrored version of the encoder.We use two-dimensional latent space for the two moons data. The architecture that is added for the decoupled prior is the same for all the experiments we present in the paper. This architecture for the affine coupling layers that connects $\mathbf{z}$ and $\mathbf{z}_0$ is shown in Figure 2.

FactorVAE [5] has a discriminator architecture, which has five fully connected layers each with 1000 hidden units. Each fully connected layer is followed by a leaky ReLU activation of negative slope of 0.2. This discriminator architecture is the same for all experiments, except for the changing input size (*i.e.,* the latent dimension L).

The learning rate for all the experiments was set to be 10^{-4}, and the batch size for MNIST was 100, SVHN and CelebA were 50. We execute all the experiments for 100,000 iterations (100,000/B epochs where B is the batch size). No other pre-processing was performed while conducting these experiments. The regularization specific hyperparameters are mentioned in Table 1. These hyperparameters were kept the same for all datasets.

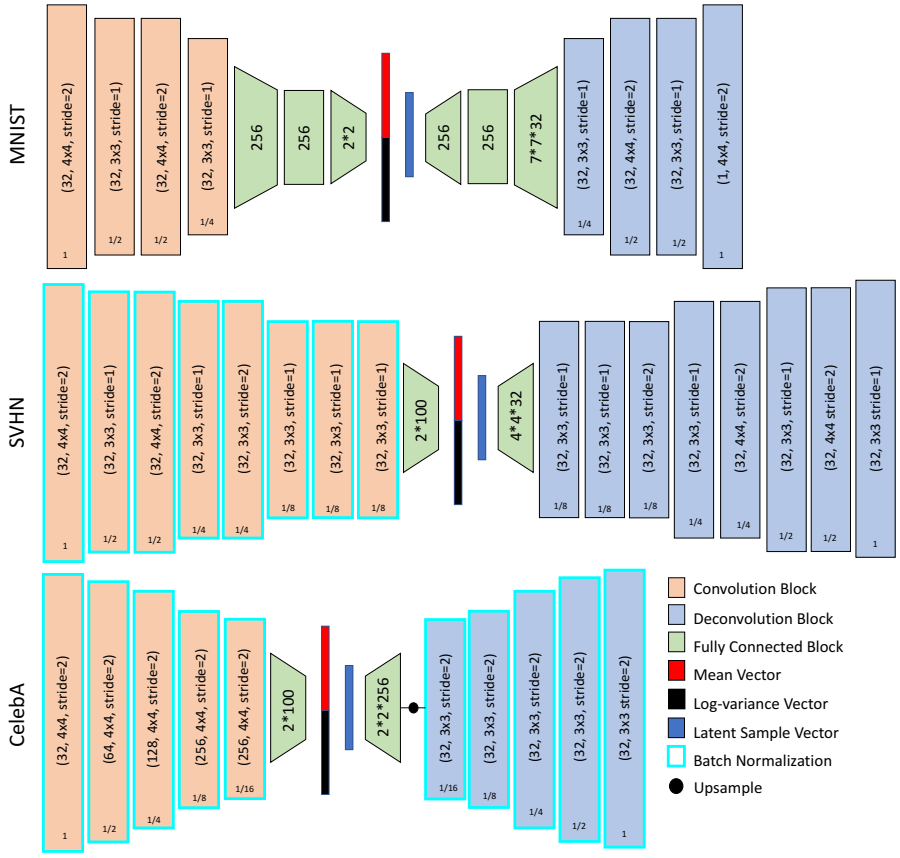

Figure 1. Architecture description for different datasets. This figure shows the VAE architecture for MNIST, SVHN, and CelebA datasets. This architecture is kept the same for all the regularizations with and without the decoupled prior.

Table 1. Table representing hyperparameters for individual regularizations. These hyperparameters were set to be the same with and without the decoupled prior.

Methods	**Parameters**
β-VAE-H [3]	$\beta = 4$
β-VAE-B [4]	$\gamma = 15$, $C_{max} = 25$, $C_{stop} = 100000$
β-TC-VAE [6]	$\alpha = 1$, $\beta = 4$, $\gamma = 15$
FactorVAE [5]	$\gamma = 1000$
InfoVAE [7]	$\alpha = 0$, $\lambda = 1000$

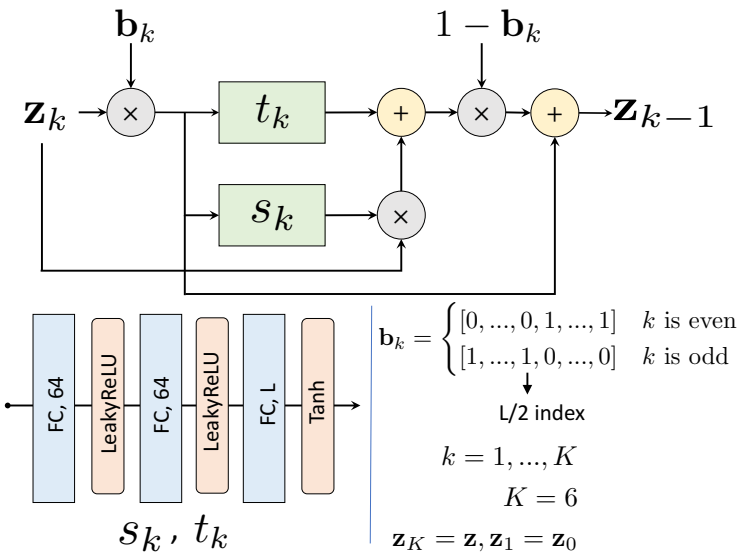

Figure 2. Architecture details of the decoupled prior. Architecture description of individual affine coupling blocks and the details of binary masks. The top shows the structure of an individual affine coupling layer representing the function $g_\eta^{(k)}(\mathbf{z}_k)$. Sndividual scaling s_k and translation t_k functions have the same architecture for all the block (bottom-left). Bottom-right shows the binary masks and number of affine coupling layers.

4 Disentanglement Results

In order to showcase that introduction of decoupled prior does not adversely affect the representation learning performance we evaluate for the disentanglement metric proposed in FactorVAE paper [5] which does not require supervised factors to evaluate. In Table 2 we compute for this metric for four different regularizers which promote disentanglement on MNIST data trained with and without decouple prior for *same number of epochs*. Even without adjusting for hyper-parameters the difference between the metrics evaluated with and without prior are negligible, giving further evidence for the proposed hypothesis.

Table 2. Disentanglement results on MNIST, with and without the VAE. The numbers represent the ability to learn disentangled features, 1 representing perfect disentanglement. We notice that even with slight drop the numbers with decoupled prior remain very close to that of without it, showcasing that introduction of prior does not adversely affect the representation learning aspect.

Method	**Without Decoupled Prior**	**With Decoupled Prior**
β-VAE-H	0.87	0.85
β-VAE-B	0.67	0.61
β-VAE-TC	0.98	0.96
FactorVAE	0.98	0.99

5 Generation Results

In this section we provide generation results on MNIST and Celeb-A datasets using the decoupled prior with different regularizers, additionally we also generate the low-posterior samples explained in Section 4.1. For MNIST these results are showcased in Figure 3. For Celeb-A we present these results for each of the regularizer in three separate images in Figures 4, 5 and 6. The latent traversal results on CelebA are shown in Figure 7.

References

1. Petersen, K.B., Pedersen, M.S.: The matrix cookbook (2012) Version 20121115.
2. Dinh, L., Sohl-Dickstein, J., Bengio, S.: Density estimation using real nvp. ICLR (2017)
3. Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., Lerchner, A.: beta-vae: Learning basic visual concepts with a constrained variational framework. ICLR **2** (2017) 6
4. Burgess, C.P., Higgins, I., Pal, A., Matthey, L., Watters, N., Desjardins, G., Lerchner, A.: Understanding disentangling in beta-vae. arXiv preprint arXiv:1804.03599 (2018)
5. Kim, H., Mnih, A.: Disentangling by factorising. In: International Conference on Machine Learning. (2018) 2654–2663

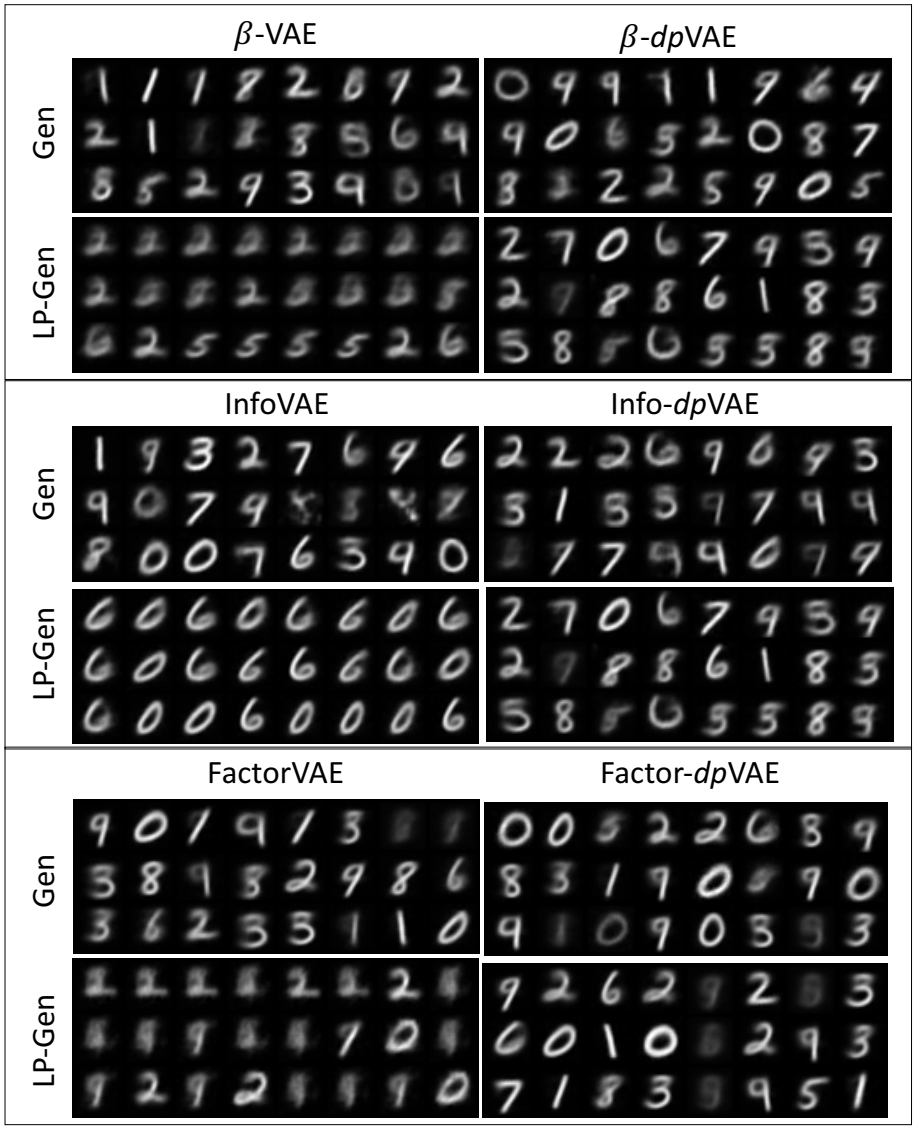

Figure 3. Genrative Results for MNIST data, on three different regularizers with and without the decoupled prior. The *Gen* tag means these are generated via standard sampling and *LP-Gen* means they are generated using the low-posterior samples.

6. Chen, T.Q., Li, X., Grosse, R.B., Duvenaud, D.K.: Isolating sources of disentanglement in variational autoencoders. In: Advances in Neural Information Processing Systems. (2018) 2610–2620
7. Zhao, S., Song, J., Ermon, S.: Infovae: Balancing learning and inference in variational autoencoders. In: Proceedings of the AAAI Conference on Artificial Intelligence. Volume 33. (2019) 5885–5892

Figure 4. Celeb-A generation results using β-VAE-B
, with and without the decoupled prior. *Gen*: generated via standard sampling, *LP-Gen*: generated via low-posterior sampling.

8. Maas, A.L., Hannun, A.Y., Ng, A.Y.: Rectifier nonlinearities improve neural network acoustic models. In: Proc. icml. Volume 30. (2013) 3

Figure 5. Celeb-A generation results using InfoVAE
, with and without the decoupled prior. *Gen*: generated via standard sampling, *LP-Gen*: generated via low-posterior sampling.

Figure 6. Celeb-A generation results using FactorVAE
, with and without the decoupled prior. *Gen*: generated via standard sampling, *LP-Gen*: generated via low-posterior sampling.

Figure 7. Latent traversals on CelebA: (a) with and (b) without decoupled prior. We notice for InfoVAE, the green boxes highlight the faces with unrealistic deformations, such as excessive teeth in the second row and noisy images in the fourth row. In comparison, Info-*dp*VAE results in a more smooth traversal. We see similar effects from β-VAE, with odd shadowing due to hair and face rotation, highlighted in blue boxes. Again, β-*dp*VAE makes these transitions smooth and seamless.