DIME: DETERMINISTIC INFORMATION MAXIMIZING AUTOENCODER

Alokendu Mazumder[1], **Chirag Garg**[2]*, **Tirthajit Baruah**[1], **Punit Rathore**[1]
[1]IISc Bengaluru, [2] NIT Raipur
{alokendum,tirthajitb,prathore}@iisc.ac.in

ABSTRACT

Variational autoencoders (VAEs) offer a theoretically sound and popular framework for deep generative models. However, learning a VAE from data presents unresolved theoretical questions and significant practical challenges. **(i)** It has been observed that the learned decoder distribution tends to be the same for all points in the latent space, implying that the latent space is not dependent on data space. This results in a poor latent representation of data. **(ii)** Additionally, due to the stochastic nature of VAE's decoder, it tends to produce blurry images that do not align well with the real data distribution, resulting in high FID scores. In this work, we propose a deterministic approach that addresses the limitations of traditional VAEs by learning a more informative latent space. Our method leverages a *von-Mises Fisher* (vMF) family-based kernel to regularize hyperspherical latent spaces in simple deterministic autoencoders. This regularization can be interpreted as maximizing the mutual information between the data and the latent space, leading to a more informative representation. We investigate how this regularization can create a better and more meaningful latent space than traditional VAE. In a rigorous empirical study, we show that our proposed model can generate samples that are comparable to, or better than, those of VAEs and other state-of-the-art autoencoders when applied to images as well as other challenging data such as equations.

1 INTRODUCTION

Generative models are essential in machine learning, enabling us to understand how data is generated, reason probabilistically, explore the low-dimensional manifolds of data, and create new data. As a result, these models are increasingly applied in fields like *computer vision* (CV) (Sohn et al., 2015; Brock et al., 2018) and *natural language processing* (NLP) (Bowman et al., 2015; Semeniuta et al., 2017).

VAEs (Kingma & Welling, 2013; Rezende et al., 2014) reformulate the task of learning representations for high-dimensional data as a variational inference problem. This involves optimizing an objective (ELBO objective) that balances the quality of autoencoded samples using a stochastic encoder-decoder pair while encouraging the latent space to adhere to a fixed prior distribution. Since their introduction, VAEs have become a widely used framework for generative modelling, often outperforming *generative adversarial networks* (GANs) (Goodfellow et al., 2014) and providing more efficient sampling than autoregressive models (Larochelle & Murray, 2011; Germain et al., 2015). There are many popular approaches to enhance the performance of VAEs for particular settings. *Adversarial autoencoders* (AAE) (Makhzani et al., 2015), and *adversarial regularized autoencoders* (ARAE) (Zhao et al., 2018) are VAE-based methods proposed to leverage adversarial learning (Goodfellow et al., 2014). Indeed, both VAE and AAE have similar goals but utilize different approaches to match the posterior with the prior. β-VAE (Higgins et al., 2017) is another variant for VAE, where the regularizer (the KL divergence between the amortized inference distribution and prior) is amplified by β. Using this, inferred high-level abstractions become more disentangled where the effect of varying each latent code is transparent.

However, VAE and its variants face both practical and theoretical challenges, including a trade-off between sample quality and reconstruction accuracy. This is often due to the use of overly simplistic prior distributions (Tomczak & Welling, 2018; Dai & Wipf, 2019) or the over-regularization from the KL divergence term in the VAE objective (Tolstikhin et al., 2017). This issue of *over-regularization* leads VAE to generate blurry images.

A significant issue is the risk of *posterior collapse*, where the latent space becomes decoupled from the input data. It generally happens when the conditional distribution of the VAE's decoder is highly expressive (Chen et al., 2020; Zhao

*Work done as an intern at IISc Bengaluru, India.

et al., 2017; Van Den Oord et al., 2017). In such scenarios, the model predominantly relies on a single component of the conditional distribution to represent the data, thereby neglecting the latent variables and missing out on the benefits of mixture modelling offered by the VAE. This undermines one of the main goals of unsupervised learning, which is to develop meaningful latent representations. Chen et al. (2016) have suggested some remedies for this issue by restricting the capacity of the conditional distribution, but this approach involves manual adjustments and is often tailored to specific problems and desired feature extraction. Later to tackle this issue, Zhao et al. (2019) propose a reformulation of the ELBO objective of VAE which maximizes the mutual information between data and latent space. Thus, Info-VAE effectively becomes a mixture of AAE and β-VAE. Using Info-VAE, the best results are achieved once the adversarial learning in AAE is replaced by the maximum-mean discrepancy. Also, Info-VAE is limited in the choice of *information preference*; see **Appendix** A.5 for more details. Another approach to optimize the mutual information is by adversarially training a critique network, which minimises a lower bound of mutual information by formulating a dual of the KL-divergence. This framework, called InfoMax VAE (Rezaabad & Vishwanath, 2020), has a limitation: the tightness of the bound cannot be guaranteed. Additionally, training the critique network alongside the encoder-decoder pair is complex and requires extra caution.

In this paper, we propose a novel solution to tackle *over-regularization* (problem of generating blurry images) and *posterior collapse* (problem of uninformative latent space). We introduce a vMF family-based regularizer in a simple deterministic autoencoder, which explicitly maximizes the mutual information between the data and the latent space. Our model, being fully deterministic, avoids generating blurry images. To sample from a deterministic decoder, we use ex-posterior density sampling (Ghosh et al., 2019b). The regularizer theoretically ensures that mutual information is preserved between the data and the latent space, effectively mitigating posterior collapse and enabling the generation of higher-quality data compared to traditional VAEs

Our model resembles InfoVAE (Zhao et al., 2019) and InfoMax VAE (Rezaabad & Vishwanath, 2020) but takes a different approach to maximizing mutual information between data and latent space. Instead of a variational framework, we use a simple regularizer that maximizes mutual information via *kernel density estimation* (KDE) with a vMF kernel, avoiding the KL-based bounds in prior methods. This streamlined formulation simplifies training. Extensive experiments show our model outperforms or matches existing methods on tasks like image generation, interpolation, and classification. Additionally, we propose a variant effective for structured data, extending its applicability beyond image datasets.

Our contributions are as follows:

- We introduce a novel regularized deterministic autoencoder for generative modelling, named *Deterministic Information Maximizing Autoencoder* (DIME).
- We theoretically show that our regularizer explicitly maximizes the mutual information between the data and latent space, resulting in a more informative latent space than VAEs.
- We theoretically prove that our regularizer provides an upper bound on the logarithm of the *maximum mean discrepancy* (MMD) between the hyperspherical uniform distribution and the encoder's output distribution, connecting it conceptually with WAE.
- We conduct a rigorous empirical evaluation, comparing DIME with VAEs and several baselines on standard image datasets and challenging structured datasets.

2 VARIATIONAL AUTOENCODERS

To be formal, let $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N\}$ denote the *i.i.d* set of observable data points drawn from a distribution $\mathcal{P}$ and $\mathcal{Z} = \{\mathbf{z}_1, \mathbf{z}_2, \cdots, \mathbf{z}_N\}$ the set of desired latent vectors, where $\mathbf{x}_i \in \mathbb{R}^d$ and $\mathbf{z}_i \in \mathbb{R}^l (l \leq d)$. Let $p_\theta(\mathbf{x}|\mathbf{z})$ denote the likelihood of generated sample conditioned on latent variable $\mathbf{z}$ and $p(\mathbf{z})$ the prior, where θ denotes the parameter set of the decoder $f_\theta : \mathcal{Z} \to \mathcal{X} \subset \mathbb{R}^d$. In classical VAEs, the prior is usually a standard Gaussian $\mathcal{N}(\mathbf{0}, \mathbb{I}_l)$. The encoder $g_\phi : \mathcal{X} \to \mathcal{Z} \subset \mathbb{R}^l$ in VAE parameterizes the variational posterior $q_\phi(\mathbf{z}|\mathbf{x})$ in light of the lower bound of the marginal log-likelihood

$$\begin{aligned}
\log p_\theta(\mathbf{x}) &= \log \int p_\theta(\mathbf{x}|\mathbf{z}) p(\mathbf{z}) d\mathbf{z} \\
&= \log \int \frac{q_\phi(\mathbf{z}|\mathbf{x})}{q_\phi(\mathbf{z}|\mathbf{x})} p_\theta(\mathbf{x}|\mathbf{z}) p(\mathbf{z}) d\mathbf{z} \\
&\geq -\mathcal{D}[q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z})] + \mathbb{E}_{q_\phi}[p_\theta(\mathbf{x}|\mathbf{z})] \qquad (1)
\end{aligned}$$

The first term, $\mathcal{D}[q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z})]$, ensures that the encoded latent codes adhere to the prior distribution through the KL divergence. The second term, $\mathbb{E}_{q_\phi}[p_\theta(\mathbf{x}|\mathbf{z})]$, guarantees the reconstruction accuracy of the inputs. For a Gaussian

$p_\theta(\mathbf{x}|\mathbf{z})$ with a diagonal covariance matrix, $\log p_\theta(\mathbf{x}|\mathbf{z})$ simplifies to the variance-weighted squared error. *From now onwards, same symbols and notations will be used in consecutive sections.*

The lower-bound approximation of the log-likelihood offers a practical solution for VAEs but introduces new challenges. For instance, generated samples $\mathbf{x} \sim p_\theta(\mathbf{x}|\mathbf{z})$ can diverge from the true distribution $\mathcal{X}$ when sampling from the prior because the learned $q_\phi(\mathbf{z}|\mathbf{x})$ struggles to match the prior distribution accurately. Additionally, the reconstruction quality may be unsatisfactory, often resulting in blurry images due to the probabilistic nature of VAEs.

3 MOTIVATION

Although VAEs remain very popular for numerous applications ranging from image processing to language modelling, they typically suffer from challenges in enabling meaningful and useful representations $\mathbf{z}$. Indeed, under appropriate situations (where the sets θ and ϕ are defined appropriately) both inference and generative models collaborate in producing an acceptable $p_\theta(\mathbf{x}|\mathbf{z})$ and an accurate amortized inference. However, finding suitable models for inference and generative networks across different tasks and datasets is challenging - when the generative model is expressive, a vanilla VAE sacrifices log-likelihood in favour of amortized inference (Chen et al., 2016). As a consequence, we obtain poor latent space which is independent from the observed data, in fact, $q_\phi(\mathbf{z}|\mathbf{x}) = q_\phi(\mathbf{z})$. *This is undesirable because a major goal of unsupervised learning is to learn meaningful latent features that should depend on the inputs.*

To understand the origin of this discrepancy, we must return to the original problem. Particularly, a maximum likelihood technique is leveraged to minimize the bound on the KL divergence between the true data distribution $q(\mathbf{x})$ and the model's marginal distribution $p_\theta(\mathbf{x})$, $\mathcal{D}[q_\phi(\mathbf{x})||p_\theta(\mathbf{x})]$. In contrast, the quality of the latent variables only depends on $q_\phi(\mathbf{z}|\mathbf{x})$. Thus, maximum likelihood without additional constraints on the posterior is insufficient to uncover relevant and information-rich latent variables. In addition, ELBO imposes a regularizer over latent codes, $\mathcal{D}[q_\phi(\mathbf{z}|\mathbf{x})||p_(\mathbf{z})]$, where it seeks in the family set of ϕ for those solutions that minimize this KL divergence. As a result, it also reduces the usefulness of latent codes by encouraging $q_\phi(\mathbf{z}|\mathbf{x})$ to be matched to $p(\mathbf{z})$, which bears no relationship with observed data. Such an approach minimizes the upper bound of the mutual information between the representations and input data. To observe this:

$$
\begin{aligned}
\mathbb{E}_{q_\phi}\left[\mathcal{D}(q_\phi(\mathbf{z}|\mathbf{x})\|p(\mathbf{z}))\right] &= \int q_\phi(\mathbf{z},\mathbf{x}) \log \frac{q_\phi(\mathbf{z}|\mathbf{x})}{p(\mathbf{z})}\, d\mathbf{x}\, d\mathbf{z} \\
&\geq \int q_\phi(\mathbf{z},\mathbf{x}) \log \frac{q_\phi(\mathbf{z}|\mathbf{x})}{p(\mathbf{z})}\, d\mathbf{x}\, d\mathbf{z} - \mathcal{D}\big[q_\phi(\mathbf{z})\|p(\mathbf{z})\big] \\
&= \int q_\phi(\mathbf{z},\mathbf{x}) \left\{ \log \frac{q_\phi(\mathbf{z}|\mathbf{x})}{p(\mathbf{z})} - \log \frac{q_\phi(\mathbf{z})}{p(\mathbf{z})} \right\} d\mathbf{x}\, d\mathbf{z} \\
&= \int q_\phi(\mathbf{z},\mathbf{x}) \log \frac{q_\phi(\mathbf{z}|\mathbf{x})}{q_\phi(\mathbf{z})}\, d\mathbf{x}\, d\mathbf{z} \\
&= I_{q_\phi}(\mathbf{x};\mathbf{z}).
\end{aligned}
$$

The inequity arises from the fact that the KL divergence does not take negative values. Hence, as vanilla VAEs push the model to minimize the KL divergence between the variational posterior $q_\phi(\mathbf{z}|\mathbf{x})$ and prior $p(\mathbf{z})$, they also force the representations to carry less information from input data. This may potentially result in inferior learned latent representations. In practice, by employing expressive generative networks, the problem is exacerbated as the model sacrifices the inference in favour of the likelihood. Indeed, the model becomes capable of recovering data from noise, regardless of latent codes. Therefore, a vanilla VAE may not be enough to discover accurate high-level abstractions of input data.

4 DETERMINISTIC INFORMATION MAXIMIZING AUTOENCODERS

As discussed earlier, VAEs without additional constraints can be unreliable for representation learning. One reason is that their objective does not appropriately consider mutual information between latent and data space. Additionally, their stochastic nature often results in the generation of blurry images. To address these issues, we propose a new deterministic autoencoder that not only maximizes mutual information but also produces sharp images compared to VAEs due to its deterministic nature. We call this model the *Deterministic Information Maximizing Autoencoder* (DIME). DIME effectively mitigates the aforementioned issues by explicitly maximizing the mutual information between representations and data within a simple deterministic framework.

Let $g_\phi : \mathcal{X} \to \mathcal{Z} \subset \mathbb{R}^l$ and $f_\theta : \mathcal{Z} \to \mathcal{X} \subset \mathbb{R}^d$ denote the encoder and decoder of a simple deterministic autoencoder respectively. Let $\mathbf{z} = g_\phi(\mathbf{x}) \in \mathcal{S}^{l-1}$ be the normalized latent vector $(i.e, \|\mathbf{z}\| = 1)$ which is fed into the decoder

f_θ. A vanilla deterministic autoencoder optimizes the *mean square error* (MSE) based reconstruction loss $\mathcal{L}_{vanilla} = \mathbb{E}_{\mathbf{x}\sim\mathcal{P}}\|\mathbf{x} - f_\theta(g_\phi(\mathbf{x}))\|^2$ without any constrain over its latent space. We propose a regularizer that explicitly maximizes mutual information between data and latent space via a vMF *kernel density estimate* (KDE). The following theorem inspires our proposed regularizer:

Theorem 1. *(Ahmad & Lin, 1976) Let p_{data} be uniform over finite samples $\{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N\}$ (e.g., a collected dataset). The redistribution entropy estimator of $g_\phi(\mathbf{x})$, where $\mathbf{x}$ follows the underlying distribution $\mathcal{P}$ that generates $\{\mathbf{x}_i\}_{i=1}^N$ via a vMF* kernel density estimate *(KDE):*

$$\mathbb{E}_{\mathbf{x}\sim p_{data}}\left[\log \mathbb{E}_{\mathbf{x}'\sim p_{data}}\left[e^{-g_\phi(\mathbf{x}')^\intercal g_\phi(\mathbf{x})/\sigma^2}\right]\right] \tag{2}$$

$$= \frac{1}{N}\sum_{i=1}^{N}\log\left(\frac{1}{N}\sum_{j=1}^{N} e^{-g_\phi(\mathbf{x}_i)^\intercal g_\phi(\mathbf{x}_j)/\sigma^2}\right) \tag{3}$$

$$= \frac{1}{N}\sum_{i=1}^{N}\log \hat{p}_{vMF}(g_\phi(\mathbf{x}_i)) + \log Z_{vMF} \tag{4}$$

$$\triangleq -\hat{H}(g_\phi(\mathbf{x})) + \log Z_{vMF}, \quad \mathbf{x}\sim\mathcal{P},$$

$$\triangleq -\hat{I}(\mathbf{x}; g_\phi(\mathbf{x})) + \log Z_{vMF}, \quad \mathbf{x}\sim\mathcal{P},$$

where

- *$\hat{p}_{vMF}$ is the KDE based on samples $\{g_\phi(\mathbf{x}_i)\}_{i=1}^N$, using a vMF kernel,*
- *Z_{vMF} is the normalization constant for vMF distribution with variance σ^2,*
- *$\hat{H}$ denotes the redistribution entropy estimator,*
- *$\hat{I}$ denotes the mutual information estimator based on $\hat{H}$, since $g_\phi(.)$ is a deterministic function.*

Proof. Kindly refer to the original paper by Ahmad & Lin (1976) for detailed proof. □

We proposed regularizing the latent space to ensure that the mutual information is maximizes between data and latent space. To achieve this, we use Eq. (4) as a regularizer along with the reconstruction loss. We define the regularized loss as the logarithm of the average pairwise Gaussian potential:

$$\mathcal{L}_{DIME}(\phi,\theta) = \underbrace{\mathbb{E}_{\mathbf{x}\sim\mathcal{P}}\|\mathbf{x} - f_\theta(g_\phi(\mathbf{x}))\|^2}_{\mathcal{L}_{DIME}^{vanilla}} + \lambda \underbrace{\log \mathbb{E}_{\mathbf{x},\mathbf{y}\sim\mathcal{P}}\left[e^{-\gamma\|g_\phi(\mathbf{x}) - g_\phi(\mathbf{y})\|^2}\right]}_{\mathcal{L}_{DIME}^{info}}, \gamma > 0 \tag{5}$$

We will now minimize the regularized loss given by Eq. (5) using *stocahstic gradient descent* (SGD) (Gower et al., 2019) optimizer in batch form. Let $\mathcal{B}$ denote a mini-batch of the training data. At each SGD iteration, a mini-batch of $|\mathcal{B}|$ points are sampled from the training set $\{\mathbf{x}_i\}_{i=1}^{|\mathcal{B}|}$, and the regularizer loss of $\mathcal{L}_{DIME}$ is computed as follows:

- $\mathcal{L}_{DIME}^{info}$: The mini-batch regularizer loss (KDE-based mutual information) is computed by:

$$\log\left(\frac{1}{|\mathcal{B}|(|\mathcal{B}|-1)}\sum_{i=1}^{|\mathcal{B}|}\sum_{j\neq i} e^{-\gamma\|g_\phi(\mathbf{x}_i) - g_\phi(\mathbf{x}_j)\|^2}\right) \tag{6}$$

The average pairwise Gaussian potential is closely associated with the redistribution entropy estimator and KDE-based mutual information estimator.

Our regularizer, as defined in Eq. (6), is not exactly the same as the mutual information estimator defined in Eq. (3), as the position of the $\log$ term differs[1]. Next, demonstrate that the minimizer of both are identical.

[1]Our regularizer in Eq. (6) applies MSE on exponentiated terms, but since $\|g_\phi(.)\| = 1$, it simplifies to an inner product, akin to Eq. (3), with an extra e^2 factor and a shifted $\log$ term.

Analysis of $\mathcal{L}_{DIME}^{info}$: Before starting with the analysis, for simplicity, we define the following notations:

Definition 1. *Let $\mathcal{M}(\mathcal{S}^{l-1})$ be the set of Borel probability measures on $\mathcal{S}^{l-1}$.*

Definition 2. *$\forall \mu \in \mathcal{M}(\mathcal{S}^{l-1}), \mu \in \mathcal{S}^{l-1}$, we define the continuous and Borel measurable function*

$$U_\mu(u) \triangleq \int_{\mathcal{S}^{l-1}} e^{u^\intercal v/\tau} d\mu(v). \tag{7}$$

with its range bounded in $[e^{-1/\tau}, e^{1/\tau}]$.

Eq.(2) can be equivalently written as

$$\mathbb{E}_{\mathbf{x}\sim p_{\text{data}}}\left[\log \mathbb{E}_{\mathbf{x}'\sim p_{\text{data}}}\left[e^{-g_\phi(\mathbf{x}')^\intercal g_\phi(\mathbf{x})/\sigma^2}\right]\right] = \mathbb{E}_{\mathbf{x}\sim p_{\text{data}}}\left[\log U_{p_{data}\circ g_\phi^{-1}} g_\phi(\mathbf{x})\right]$$

where $p_{data} \circ g_{phi}^{-1} \in \mathcal{M}(\mathcal{S}^{l-1})$ is the probability measure of features, i.e, the pushforward measure of p_{data} via g_ϕ.

We now present the theorem that establishes the equivalence of the minimizers for Eq. (6) and Eq. (3). This result is based on the derivation by Wang & Isola (2020).

Theorem 2. *Consider the following relaxed problem, where the minimization is taken over $\mathcal{M}(\mathcal{S}^{l-1})$, all possible Borel probability measures on the hypersphere $\mathcal{S}^{l-1}$:*

$$\min_{\mu\in\mathcal{M}(\mathcal{S}^{l-1})} \int_{\mathcal{S}^{l-1}} \log U_\mu(v) d\mu(u). \tag{8}$$

The minimizer of Eq.(8) is unique and is identical to the minimizer of the following problem.

$$\min_{\mu\in\mathcal{M}(\mathcal{S}^{l-1})} \log \int_{\mathcal{S}^{l-1}} U_\mu(v) d\mu(u)$$

Proof. Kindly refer to **Theorem 1** of the paper by Wang & Isola (2020). □

Theorem 2 demonstrates that Eq. (6) and Eq. (3) share the same minimizer. Hence, pulling out the $\log$ from the outer integral[2] does not alter the solution. However, if we push the $\log$ all the way inside the inner integral, the problem becomes equivalent to minimizing the norm of the mean, i.e.,

$$\min_{\mu\in\mathcal{M}(\mathbb{S}^{l-1})} \mathbb{E}_{U\sim\mu}[U]^\top \mathbb{E}_{U\sim\mu}[U],$$

which is minimized for any distribution with the mean being the all-zeros vector $\mathbf{0}$, e.g.,

$$\frac{1}{2}\delta_u + \frac{1}{2}\delta_{-u}, \quad \forall u \in \mathcal{S}^{l-1},$$

(where δ_u is the Dirac delta distribution at u such that $\delta_u(S) = \mathbf{1}_S(u), \forall S$). Therefore, the location of the $\log$ is important.

5 Theoritical Analysis

This section offers a comprehensive theoretical analysis of our proposed regularizer. We investigate the dynamics of the latent space with our proposed regularizer and establish connections with the famous *Wasserestein autoencoder* (WAE) (Tolstikhin et al., 2017). We demonstrate that our proposed regularizer serves as a lower bound on the logarithm of the *maximum mean discrepancy* (MMD) between the hyperspherical uniform distribution and the unknown distribution of the latent space under vMF kernel, thereby establishing a connection with WAE (**Theorem** 3).

Connections with WAE: The maximum mean discrepancy is a divergence measure between two distributions $\mathcal{P}$ and $\mathcal{Q}$. In the context of WAEs, applying the encoder to the distribution of the input data (e.g. images) yields the aggregate distribution $\mathcal{Q}$ of the latent variables. One of the goals of WAE training is to make $\mathcal{Q}$ (which depends on the neural net parameters) as close as possible to some fixed target distribution $\mathcal{P}$. This is achieved by incorporating MMD between $\mathcal{P}$ and $\mathcal{Q}$ as a regularizer into the WAE objective.

[2] U_μ inherently has an integral, which is the inner integral. See Eq. (7)

The computation of the MMD requires specifying a positive-definite kernel $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, as per **Theorem** 1, we assume it to be a Gaussian RBF kernel with bandwidth σ. The population MMD can be most straight-forwardly computed via the formula (Gretton et al., 2012):

$$\text{MMD}^2(\mathcal{P}, \mathcal{Q}) = \mathbb{E}_{\mathbf{w},\mathbf{w}'\sim\mathcal{P}}[k(\mathbf{w}, \mathbf{w}')] - 2\mathbb{E}_{\mathbf{w}\sim\mathcal{P},\mathbf{z}\sim\mathcal{Q}}[k(\mathbf{w}, \mathbf{z})] + \mathbb{E}_{\mathbf{z},\mathbf{z}'\sim\mathcal{Q}}[k(\mathbf{z}, \mathbf{z}')] \tag{9}$$

We start with the expression Eq.(9) and using the samples from $\mathcal{Q}_N = \{\mathbf{z}_i\}_{i=1}^N$, we replace the last two terms by the sample average and the U-statistic respectively to obtain the unbiased estimator (Gretton et al., 2012):

$$MMD_N^2(\mathcal{P}, \mathcal{Q}) = \mathbb{E}_{\mathbf{w},\mathbf{w}'\sim\mathcal{P}}[k(\mathbf{w}, \mathbf{w}')] - \frac{2}{N}\sum_{i=1}^{N}\mathbb{E}_{\mathbf{w}\sim\mathcal{P}}[k(\mathbf{w}, \mathbf{z}_i)] + \frac{1}{N(N-1)}\sum_{i=1}^{N}\sum_{j\neq i}^{N} k(\mathbf{z}_i, \mathbf{z}_j). \tag{10}$$

Here $\mathbf{w}, \mathbf{z} \in \mathcal{S}^{l-1}$ are unit length latent vectors. We assume the distribution $\mathcal{P}$ to be a uniform distribution over $(l-1)$ dimensional sphere with density $p(\mathbf{w}) = \frac{\Gamma(l/2)}{2\pi^{l/2}}$. In practical situations, we only have access to $\mathcal{Q}$ through a sample. For example, during each step of the WAE training, the encoder will compute $\{\mathbf{z}_i\}_{i=1}^N$ corresponding to the input data and the current values of neural network parameters.

Theorem 3. *The closed form of $MMD_N^2(\mathcal{P}, \mathcal{Q})$, where $\mathcal{P}$ is a uniform distribution on $\mathcal{S}^{l-1}$ can be computed analytically (with N samples) to yield the following:*

$$\textit{MMD}_N^2(\mathcal{P}_N, \mathcal{Q}_N) = -a(\sigma, l) + \frac{1}{N(N-1)}\sum_{i=1}^{N}\sum_{j\neq i}^{N} e^{-\frac{\|\mathbf{z}_i - \mathbf{z}_j\|^2}{2\sigma^2}} \tag{11}$$

and, our proposed regularizer is related with $\textit{MMD}_N^2$ as follows:

$$\log \textit{MMD}_N^2(\mathcal{P}_N, \mathcal{Q}_N) \leq \mathcal{L}_{DIME}^{info}.$$

Here, $a(\sigma, l) = e^{-\sigma^{-2}}\Gamma(l/2)\left(2\sigma^2\right)^{\frac{l-2}{2}}\mathbf{I}_{\frac{l-2}{2}}\left(\sigma^{-2}\right)$ and $\mathbf{I}_\nu(.)$ is the modified Bessel function of the first kind.

Proof. See **Appendix** A.4. □

Our regularizer serves as the upper bound of the closed-form solution of the log of MMD when reference distribution $\mathcal{P}$ is uniform over $(l-1)$ dimensional sphere. When the latent vectors are un-normalized, one can choose $\mathcal{P}$ as a standard normal distribution over the l dimensional space (classic WAE); in this case, the inequality will be flipped.

In this work, our primary focus is on developing a novel autoencoder model that maximizes the mutual information between the data and the latent space. Hence, we refrain from delving into a detailed analysis of the tightness of the bound, leaving it as a direction for future research.

6 Related Work

Deferred to **Appendix** A.1 due to space constraint.

7 Experiments

Our experiments are designed to answer whether sample quality and latent space structure in DIME is comparable to VAEs and other autoencoders? The quality and effectiveness of a learned latent space can be evaluated by the success of downstream tasks, such as generating images, downstream classification and observing the smooth transition between different points in the data space.

7.1 Image Generation

High-quality image generation signifies successful learning in the latent space. Since the DIME operates as a deterministic model, it lacks a latent distribution from which images can directly be sampled. To overcome this limitation during the inference phase, we apply ex-post density estimation (Ghosh et al., 2019b; Jing et al., 2020; Mazumder et al., 2024) on $\mathbf{z}$. We achieve this by fitting a density estimator, $q_\psi(\mathbf{z})$, to the discrete set $\mathcal{C} := \{\mathbf{z} = g_{\phi^*}(\mathbf{x}) | \mathbf{x} \in \mathcal{X}\}$. This technique not only integrates seamlessly with the DIME framework but also offers a viable solution for other deterministic AEs like the WAE and RAE, effectively addressing the aggregated posterior mismatch without imposing additional computational demands during the training phase.

Table 1: Qualitative evaluation of sample quality for VAE, AE, state-of-the-art deterministic autoencoders (WAE and RAE) and DIME on CelebA reveals that DIME stands out by generating cleaner samples and interpolating smoothly in the latent space. DIMEE also reconstructs sharper images despite incorporating a regularization term. *Interestingly, the reconstruction quality does not suffer due to the presence of regularizer*. In contrast, VAE, being probabilistic, produces blurry images. More qualitative results on different datasets are given in **Appendix** A.6.

Selecting $q_\psi(\mathbf{z})$: The selection of $q_\psi(\mathbf{z})$ balances between being expressive enough to accurately model the diverse latent space of $\mathbf{z}$ and being simple enough to ensure generalization. For example, adopting a Dirac distribution at each latent point $\mathbf{z}$ might result in high-fidelity reconstructions of training samples but would fail to generalize well. In our experiments, we opted for simplicity by employing a 10-component *Gaussian mixture model* (GMM) and *multivariate Gaussian* (MVG or $\mathcal{N}$) noise to model the latent space in deterministic autoencoders, facilitating the generation of images. This is a widely known technique that is extensively used for generating images from deterministic encoder-decoder pairs (Ghosh et al., 2019b; Jing et al., 2020; Mazumder et al., 2024).

	MNIST		CIFAR-10		CelebA	
	$\mathcal{N}$	GMM	$\mathcal{N}$	GMM	$\mathcal{N}$	GMM
VAE	34.75	42.30	281.42	283.86	61.69	65.14
β-VAE	19.34	12.65	113.18	90.3	53.89	50.65
InfoVAE	36.71	29.44	251.59	238.43	66.65	63.90
$\mathcal{S}$-VAE	53.89	–	–	–	–	–
HVAE	26.96	27.76	274.16	261.14	59.48	51.48
VAE-LinNF	29.31	28.46	240.32	247.09	56.54	53.36
VAE-IAF	27.53	27.00	236.08	235.43	55.43	53.61
CV-VAE	33.79	17.87	94.75	86.64	48.87	49.30
2s-VAE	<u>18.81</u>	–	109.77	–	49.70	–
AE	29.79	16.78	<u>91.67</u>	85.65	69.43	62.35
WAE	22.86	12.22	111.44	<u>84.77</u>	50.59	43.12
RAE-GP	24.61	13.01	92.90	84.97	<u>48.45</u>	**38.70**
RAE-L_2	25.92	12.45	92.03	84.79	51.84	45.10
IRAME	26.58	22.31	–	–	58.98	48.96
LoRAE	19.50	<u>11.69</u>	–	–	56.29	46.43
DIME **(Ours)**	**17.97**	**10.32**	**89.01**	**83.16**	**45.30**	<u>39.36</u>

Table 2: Evaluation of all models by FID (↓, best models in bold, second best in underline). We evaluate each model by **(i)** $\mathcal{N}$: random samples generated according to the prior distribution $p(\mathbf{z})$ (isotropic Gaussian for VAE, β-VAE, HVAE, VAE-LinNF, VAE-IAF, CV-VAE, 2s-VAE, and WAE; spherical uniform distribution for $\mathcal{S}$-VAE) or by fitting a Gaussian to $q_\psi(\mathbf{z})$ (ex-post density estimation for the remaining models); **(ii)** GMM: random samples generated by fitting a mixture of 10 Gaussians in the latent space.

We trained our model using the MNIST (Deng, 2012), CIFAR-10 (Krizhevsky et al., 2009) and CelebA (Liu et al., 2015) datasets and conducted a comparison with a vast family of generative as well as deterministic autoencoders. The generative family includes VAE (Kingma & Welling, 2013), β-VAE (Higgins et al., 2017), Info-VAE (Zhao et al., 2019), $\mathcal{S}$-VAE (Davidson et al., 2018),

HVAE (Caterini et al., 2018), VAE-LinNF (Rezende & Mohamed, 2015), VAE-IAF (Kingma et al., 2016), CV-VAE (Ballé et al., 2016; Ghosh et al., 2019a) and 2s-VAE (Dai & Wipf, 2019). The deterministic family includes a simple AE, WAE (Tolstikhin et al., 2017), RAE (Ghosh et al., 2019b), IRMAE (Jing et al., 2020) and LoRAE (Mazumder et al., 2024). We set the latent space dimensions to 16, 128 and 64 for MNIST, CIFAR-10 and CelebA, respectively, across all models[3], training for 100 epochs each. We demonstrate the capability of DIME to generate high-quality images via sampling from Gaussian noise. In particular, we accomplish this by fitting (i) a 10-component GMM (to avoid overfitting) and (ii) an MVG (which we denote as $\mathcal{N}$ in **Table** 2) distribution to its latent space. After fitting the noise distribution, we proceed to sample from it and feed these samples through the decoder to generate images. We quantitatively evaluate the generating capacity of each model using the *Frechet inception distance* (FID) (Heusel et al., 2017; Parmar et al., 2022) score. Analyzing the outcomes presented in **Table** 1, a clear trend emerges: our model excels in generating images with higher visual quality when compared to the simple AE and VAE. Furthermore, its image generation quality stands on par with that of state-of-the-art autoecoders like WAE and RAE. Differing from VAE, which often produces images with blurred backgrounds due to its probabilistic nature, our model, being deterministic, avoids generating images with such blurriness. This distinction results in clearer and more defined images produced by DIME. These findings are supported by the data showcased in **Table** 2 where our model achieves the best FID score among the 14 benchmarked autoencoders (standing second for one category), highlighting the pivotal role of hyperspherical and uniform latent space achieved through explicit regularization.

7.2 Image Classification

Deferred to **Appendix** A.2 due to space constraint.

8 Grammar-DIME: Modelling Structured Inputs

We evaluate DIME for generating complex structured objects, such as arithmetic expressions, with two primary goals: **(i)** to explore how well DIME learns latent spaces for more challenging inputs that adhere to specific structural constraints, and **(ii)** to assess the benefits of replacing VAE with DIME in a state-of-the-art generative model. When experimenting with structured data like equations, we refer to our model as *Grammar*DIME (GDIME). We replicate the architectures and experimental settings of *Grammar*VAE (GVAE) (Kusner et al., 2017), which has demonstrated superiority over other generative models such as *Character*VAE (CVAE) (Gómez-Bombarelli et al., 2018). As in Kusner et al. (2017), our focus is on exploring the latent space learned by our models to generate mathematical expressions that optimize a specific downstream metric. This is achieved through *Bayesian optimization* (BO) using the $\log(1 + MSE)$ metric (lower is better) for the generated expressions compared to some ground truth points. A well-structured latent space will not only generate expressions with better scores during the BO process but will also produce syntactically valid ones, adhering to the grammar rules of the problem.

Objective	Method	Expressions
LL	GVAE	-1.320 $\pm$ 0.001
	CVAE	-1.397 $\pm$ 0.003
	GDIME	-1.261 $\pm$ 0.001
RMSE	GVAE	0.884 $\pm$ 0.002
	CVAE	0.875 $\pm$ 0.004
	GDIME	0.830 $\pm$ 0.001

Table 3: Average test root mean square error (RMSE) and test log-likelihood (LL) for the DIME, CVAE, and GVAE across 10 different splits of the data for the expressions.

Problem	Method	Frac. Valid	Avg. Score
Expression	GVAE	0.99 $\pm$ 0.01	3.47 $\pm$ 0.24
	CVAE	0.86 $\pm$ 0.06	3.75 $\pm$ 0.25
	GDIME	1.00 $\pm$ 0.00	3.16 $\pm$ 0.21

Table 4: Percentage of valid samples of equations generated by GVAE, CVAE, and GDIME and their average mean score.

Table 3 presents the results from 5 trials of *Bayesian optimization* (BO). Our GDIME outperforms CVAE and GVAE in generating mathematical expressions, achieving superior LL and RMSE scores. Additionally, as shown in **Table** 4, GDIME produces a higher number of valid samples compared to GVAE and CVAE and achieves a better average score. These findings suggest that DIME not only excels in generating image data but also demonstrates strong performance in structured data generation, highlighting its versatility and effectiveness across various datasets.

9 Conclusion

In this paper, we address the issue of representation collapse in VAEs. Our findings reveal that the conventional objective of VAEs is inadequate for learning general and useful representations. We also observe that complex generative networks tend to inhibit the model from acquiring constructive representations. To tackle these challenges, we propose a novel information-based, simple, and deterministic autoencoder that maximizes the information retained in the latent representations from the observations.

[3]Details about model architecture and hyperparameters are deferred to **Appendix** A.3 due to space constraint.

Through extensive experiments, we demonstrate that our model outperforms or matches the performance of other state-of-the-art autoencoders on both image datasets and challenging structured datasets.

ACKNOWLEDGMENTS

This work is supported by *Prime Minister's research Fellowship* (PMRF), India.

REFERENCES

Ibrahim Ahmad and Pi-Erh Lin. A nonparametric estimation of the entropy for absolutely continuous distributions (corresp.). *IEEE Transactions on Information Theory*, 22(3):372–375, 1976.

Alexander Alemi, Ben Poole, Ian Fischer, Joshua Dillon, Rif A Saurous, and Kevin Murphy. Fixing a broken elbo. In *International conference on machine learning*, pp. 159–168. PMLR, 2018.

Johannes Ballé, Valero Laparra, and Eero P Simoncelli. End-to-end optimized image compression. *arXiv preprint arXiv:1611.01704*, 2016.

Matthias Bauer and Andriy Mnih. Resampled priors for variational autoencoders. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 66–75. PMLR, 2019.

Yoshua Bengio, Li Yao, Guillaume Alain, and Pascal Vincent. Generalized denoising auto-encoders as generative models. *Advances in neural information processing systems*, 26, 2013.

David Berthelot, Colin Raffel, Aurko Roy, and Ian Goodfellow. Understanding and improving interpolation in autoencoders via an adversarial regularizer. *arXiv preprint arXiv:1807.07543*, 2018.

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.

Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018.

Anthony L Caterini, Arnaud Doucet, and Dino Sejdinovic. Hamiltonian variational auto-encoder. *Advances in Neural Information Processing Systems*, 31, 2018.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PMLR, 2020.

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.

Mary Kathryn Cowles and Bradley P Carlin. Markov chain monte carlo convergence diagnostics: a comparative review. *Journal of the American statistical Association*, 91(434):883–904, 1996.

Bin Dai and David Wipf. Diagnosing and enhancing vae models. *arXiv preprint arXiv:1903.05789*, 2019.

Tim R Davidson, Luca Falorsi, Nicola De Cao, Thomas Kipf, and Jakub M Tomczak. Hyperspherical variational auto-encoders. *arXiv preprint arXiv:1804.00891*, 2018.

Li Deng. The mnist database of handwritten digit images for machine learning research [best of the web]. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012. doi: 10.1109/MSP.2012.2211477.

Adji B Dieng, Yoon Kim, Alexander M Rush, and David M Blei. Avoiding latent variable collapse with generative skip models. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 2397–2405. PMLR, 2019.

Mathieu Germain, Karol Gregor, Iain Murray, and Hugo Larochelle. Made: Masked autoencoder for distribution estimation. In *International conference on machine learning*, pp. 881–889. PMLR, 2015.

Partha Ghosh, Arpan Losalka, and Michael J Black. Resisting adversarial attacks using gaussian mixture variational autoencoders. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pp. 541–548, 2019a.

Partha Ghosh, Mehdi SM Sajjadi, Antonio Vergari, Michael Black, and Bernhard Schölkopf. From variational to deterministic autoencoders. *arXiv preprint arXiv:1903.12436*, 2019b.

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.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

Robert Mansel Gower, Nicolas Loizou, Xun Qian, Alibek Sailanbayev, Egor Shulgin, and Peter Richtárik. Sgd: General analysis and improved rates. In *International conference on machine learning*, pp. 5200–5209. PMLR, 2019.

Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

Irina Higgins, Loic Matthey, Arka Pal, Christopher P Burgess, Xavier Glorot, Matthew M Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. *ICLR (Poster)*, 3, 2017.

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.

Li Jing, Jure Zbontar, et al. Implicit rank-minimizing autoencoder. *Advances in Neural Information Processing Systems*, 33: 14736–14746, 2020.

Hyunjik Kim and Andriy Mnih. Disentangling by factorising. In *International conference on machine learning*, pp. 2649–2658. PMLR, 2018.

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

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.

Matt J Kusner, Brooks Paige, and José Miguel Hernández-Lobato. Grammar variational autoencoder. In *International conference on machine learning*, pp. 1945–1954. PMLR, 2017.

Hugo Larochelle and Iain Murray. The neural autoregressive distribution estimator. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pp. 29–37. JMLR Workshop and Conference Proceedings, 2011.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

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

Kanti V Mardia and Peter E Jupp. *Directional statistics*. John Wiley & Sons, 2009.

Alokendu Mazumder, Tirthajit Baruah, Bhartendu Kumar, Rishab Sharma, Vishwajeet Pattanaik, and Punit Rathore. Learning low-rank latent spaces with simple deterministic autoencoder: Theoretical and empirical insights. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 2851–2860, 2024.

Gaurav Parmar, Richard Zhang, and Jun-Yan Zhu. On aliased resizing and surprising subtleties in gan evaluation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11410–11420, 2022.

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, 2019.

Ali Lotfi Rezaabad and Sriram Vishwanath. Learning representations by maximizing mutual information in variational autoencoders. In *2020 IEEE International Symposium on Information Theory (ISIT)*, pp. 2729–2734. IEEE, 2020.

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.

Salah Rifai, Pascal Vincent, Xavier Muller, Xavier Glorot, and Yoshua Bengio. Contractive auto-encoders: Explicit invariance during feature extraction. In *Proceedings of the 28th international conference on international conference on machine learning*, pp. 833–840, 2011.

Salah Rifai, Yoshua Bengio, Yann Dauphin, and Pascal Vincent. A generative process for sampling contractive auto-encoders. *arXiv preprint arXiv:1206.6434*, 2012.

Stanislau Semeniuta, Aliaksei Severyn, and Erhardt Barth. A hybrid convolutional variational autoencoder for text generation. *arXiv preprint arXiv:1702.02390*, 2017.

Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. *Advances in neural information processing systems*, 28, 2015.

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.

Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.

Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning*, pp. 1096–1103, 2008.

Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International conference on machine learning*, pp. 9929–9939. PMLR, 2020.

Deli Zhao, Jiapeng Zhu, and Bo Zhang. Latent variables on spheres for sampling and inference.

Junbo Zhao, Yoon Kim, Kelly Zhang, Alexander Rush, and Yann LeCun. Adversarially regularized autoencoders. In *International conference on machine learning*, pp. 5902–5911. PMLR, 2018.

Shengjia Zhao, Jiaming Song, and Stefano Ermon. Towards deeper understanding of variational autoencoding models. *arXiv preprint arXiv:1702.08658*, 2017.

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 RELATED WORK

Autoencoder literature: Numerous studies have explored diagnosing the VAE framework by analyzing its objective function components (Hoffman & Johnson, 2016; Zhao et al., 2017; Alemi et al., 2018) and suggesting enhancements to overcome optimization challenges (Rezende et al., 2014; Dai & Wipf, 2019). While spherical inference (Zhao et al.) has been proposed as an alternative to variational inference, it typically performs well only in low-dimensional latent spaces. In contrast, we suggest that a simpler deterministic approach, such as DIME, can be just as effective for generative modelling. Deterministic denoising (Vincent et al., 2008) and *contractive autoencoders* (CAEs) (Rifai et al., 2011) have been investigated for their ability to capture smooth data manifolds. Various heuristic methods have attempted to imbue these models with generative abilities, including through MCMC schemes (Rifai et al., 2012; Bengio et al., 2013). However, these methods present challenges related to convergence diagnosis, require extensive tuning (Cowles & Carlin, 1996), and have not been effectively scaled beyond MNIST, which led to their replacement by VAEs.

Addressing the mismatch between the aggregated posterior and the prior $p(\mathbf{z})$ often involves making the prior more expressive (Kingma & Welling, 2013; Bauer & Mnih, 2019), which alters the VAE objective and demands significant additional computational resources. Using another VAE to estimate the latent space of the original VAE (Dai & Wipf, 2019) reintroduces many optimization challenges and it is much more resource-intensive compared to simply fitting a straightforward $q_\psi(\mathbf{z})$ post-training. *Adversarial autoencoders* (AAEs) (Makhzani et al., 2015) add a discriminator to the deterministic encoder-decoder pair, producing sharper samples but at the cost of increased computational complexity and potential instability due to adversarial training. *Wasserstein autoencoders* (WAEs) (Tolstikhin et al., 2017) generalize AAEs by framing autoencoding as an *optimal transport* (OT) problem. Both stochastic and deterministic models can be trained by minimizing a relaxed OT cost function, using either an adversarial loss term or the maximum mean discrepancy score between $p(\mathbf{z})$ and $q_\phi(\mathbf{x})$ as a regularizer instead of KL-divergence. The most effective VAE architectures for images and audio to date are variations of the *vector quantized variational autoencoders* (VQ-VAE) (Van Den Oord et al., 2017; Razavi et al., 2019). Despite the name, VQ-VAEs are neither stochastic nor variational; rather, they are deterministic autoencoders. However, VQ-VAEs require complex discrete autoregressive density estimators and a training loss that is non-differentiable due to the quantization of $\mathbf{z}$.

Bridging the gap between stochastic and deterministic encoders: A recent shift in the field has seen a growing preference for deterministic autoencoders over generative ones for data generation. These deterministic models are easier to train, do not have variational inference, and avoid producing blurry images. Previous works have employed regularization techniques such as imposing low-rank constraints in the latent space (Jing et al., 2020; Mazumder et al., 2024) and adding Gaussian noise to the input of deterministic autoencoders (Ghosh et al., 2019b). Although these models have demonstrated significant success in generating high-quality images, they fail to ensure that the mutual information between the latent space and the data space is preserved.

Information theoretic literature: Recent research has focused on enhancing VAE by maximizing mutual information. For instance, Dieng et al. (2019) introduced skip connections from the latent variables to the VAE output, implicitly reinforcing the dependency between latent representations and observations. Similarly, Hoffman & Johnson (2016) proposes directly optimizing the mutual information between latent representations and input data by incorporating a mutual information term into the VAE objective. This approach relies on Monte Carlo estimation of $q_\psi(\mathbf{z})$, but the computational burden can hinder performance improvements (Kim & Mnih, 2018). To address this challenge, InfoVAE (Zhao et al., 2019) circumvents explicit mutual information estimation by reformulating the objective function. Instead of direct computation, it minimizes the MMD or KL divergence between the marginal distribution of the inference network and the prior, thereby increasing the mutual information implicitly. This formulation aligns InfoVAE with elements of both AAE and β-VAEs. The best results with InfoVAE are obtained when adversarial learning in AAE is substituted with MMD. However, InfoVAE has limitations in controlling the degree of information preservation (see **Appendix** A.5 for more details). Similar to InfoVAE, Rezaabad & Vishwanath (2020) explicitly maximizes mutual information in a VAE framework using the dual form of KL divergence. This requires training an additional critic network, leading to unstable training. Moreover, as an extension of the VAE framework, it inherits the common issue of generating blurry images. In contrast, our approach explicitly estimates and maximizes mutual information using a KDE-based classical method. This approach enables capturing richer latent representations than InfoVAE, which is evident from **Table** 1. Furthermore, due to its deterministic nature, DIME produces sharper images than traditional VAE variants, resulting in consistently superior performance across a range of models and datasets.

A.2 EXPERIMENTS: IMAGE CLASSIFICATION

To evaluate the quality of learned latent representations, we perform a classification task using a single-layer classifier, as well-disentangled representations enable strong linear separability (Berthelot et al., 2018). Classifiers are trained on MNIST and CIFAR-10 embeddings with the same train/val/test splits as the autoencoder training. Each model is evaluated over 20 runs, with test performance reported in **Table** 5. Deterministic AEs outperform variational methods, as the latter enforce continuity, bringing class representations closer. Notably, DIME achieves the highest accuracy, highlighting its ability to learn discriminative latent spaces. Models with flexible priors generally perform better.

	VAMP	VAE-LinAF	VAE-IAF	$\mathcal{S}$-VAE	Info-VAE	HVAE	β-VAE	RAE-L_2	RAE-GP	WAE-MMD	VAE	AE	DIME
MNIST	<u>93.77</u>	85.95	90.59	86.97	73.63	90.70	87.44	92.35	90.55	88.21	90.41	91.48	**94.50**
CIFAR-10	48.97	–	48.75	–	32.49	51.02	49.14	54.02	<u>54.05</u>	53.84	51.34	53.41	**55.83**

Table 5: Classification accuracy for MNIST and CIFAR-10 using different autoencoders.

A.3 Model Architecture and Hyperparameter Details

Datasets	MNIST	Cifar-10	CelebA
Encoder	$x \in \mathbb{R}^{32\times32\times1}$ $\rightarrow Conv_{32} \rightarrow ReLU$ $\rightarrow Conv_{64} \rightarrow ReLU$ $\rightarrow Conv_{128} \rightarrow ReLU$ $\rightarrow Conv_{256} \rightarrow ReLU$ Flatten 1024 $\rightarrow FC_{128} \rightarrow z \in \mathbb{R}^{16}$	$x \in \mathbb{R}^{64\times64\times3}$ $\rightarrow Conv_{128} \rightarrow BN + ReLU$ $\rightarrow Conv_{256} \rightarrow BN + ReLU$ $\rightarrow Conv_{512} \rightarrow BN + ReLU$ $\rightarrow Conv_{1024} \rightarrow BN + ReLU$ Flatten 16,384 $\rightarrow FC_{4096} \rightarrow z \in \mathbb{R}^{128}$	$x \in \mathbb{R}^{64\times64\times3}$ $\rightarrow Conv_{128} \rightarrow BN + ReLU$ $\rightarrow Conv_{256} \rightarrow BN + ReLU$ $\rightarrow Conv_{512} \rightarrow BN + ReLU$ $\rightarrow Conv_{1024} \rightarrow BN + ReLU$ Flatten 16,384 $\rightarrow FC_{4096} \rightarrow z \in \mathbb{R}^{64}$
Decoder	$z \in \mathbb{R}^{16}$ FC_{8096} Reshape to $8 \times 8 \times 128$ $\rightarrow ConvT_{64} \rightarrow ReLU$ $\rightarrow ConvT_{32} \rightarrow ReLU$ $\rightarrow ConvT_{3} \rightarrow Tanh$ $\hat{x} \in \mathbb{R}^{32\times32\times1}$	$z \in \mathbb{R}^{128}$ $FC_{65536} \rightarrow ReLU$ Reshape to $8 \times 8 \times 1024$ $\rightarrow ConvT_{1024} \rightarrow BN + ReLU$ $\rightarrow ConvT_{512} \rightarrow BN + ReLU$ $\rightarrow ConvT_{256} \rightarrow BN + ReLU$ $\rightarrow ConvT_{3} \rightarrow Sigmoid$ $\hat{x} \in \mathbb{R}^{64\times64\times3}$	$z \in \mathbb{R}^{64}$ $FC_{65536} \rightarrow ReLU$ Reshape to $8 \times 8 \times 1024$ $\rightarrow ConvT_{1024} \rightarrow BN + ReLU$ $\rightarrow ConvT_{512} \rightarrow BN + ReLU$ $\rightarrow ConvT_{256} \rightarrow BN + ReLU$ $\rightarrow ConvT_{128} \rightarrow BN + ReLU$ $\rightarrow ConvT_{3} \rightarrow Sigmoid$ $\hat{x} \in \mathbb{R}^{64\times64\times3}$

Table 6: Architecture of encoder and decoder for MNIST, Cifar and CelebA dataset. In our model (EIMAE), the latent vectors are normalized before feeding into the decoder.

Dataset	MNIST	Cifar-10	CelebA
Batch Size	100	100	100
Epochs	100	100	100
Training Examples	60,000	50,000	16,2079
Test Examples	10,000	10,000	20,000
Learning Rate	5×10^{-3}	5×10^{-3}	5×10^{-3}
t	5×10^{-3}	5×10^{-3}	5×10^{-3}
λ	5×10^{-4}	10^{-2}	10^{-2}

Table 7: The hyperparameters for each experiment are elaborated in the following table. The determination of the number of epochs was guided by the aim of attaining a stage of converged reconstruction error.

A.4 Proofs

Before proving **Theorem** 3, we first introduce the following lemma.

Lemma 1. *Let* $\mathbf{x}$ *and* $\mathbf{z}$ *be independent random vectors uniformly distributed on the unit sphere* $\mathcal{S}^{l-1}$*. Then the inner product*

$$u = \mathbf{x}^\intercal \mathbf{z}$$

has probability density function

$$p(u) = \frac{\Gamma\left(\frac{l}{2}\right)}{\sqrt{\pi}\,\Gamma\left(\frac{l-1}{2}\right)} \left(1-u^2\right)^{\frac{l-3}{2}}, \quad u \in [-1,1].$$

Proof. Since the joint distribution of $\mathbf{x}$ and $\mathbf{z}$ is invariant under rotations, the distribution of the scalar $u = \mathbf{x}^\intercal \mathbf{z}$ depends only on the angle between $\mathbf{x}$ and $\mathbf{z}$. (In a rigorous argument one does not fix $\mathbf{z}$, but by rotational invariance one may show that the marginal

distribution of u is identical to that obtained by first rotating the coordinate system so that $\mathbf{z} = (1, 0, \ldots, 0)$; the final result is independent of this choice.)

In that coordinate system we have

$$u = \mathbf{x}^\intercal \mathbf{z} = x_1,$$

where $\mathbf{x} = (x_1, x_2, \ldots, x_l)$ is uniformly distributed on $\mathcal{S}^{l-1}$. It is a standard fact given by Mardia & Jupp (2009) that the marginal density of the first coordinate x_1 is given by

$$p(x_1) = C_l (1 - x_1^2)^{\frac{l-3}{2}}, \quad x_1 \in [-1, 1],$$

where the normalization constant C_l is determined by

$$\int_{-1}^{1} C_l (1 - x_1^2)^{\frac{l-3}{2}} \, dx_1 = 1.$$

Using the standard Beta integral, one can show that

$$C_l = \frac{\Gamma\left(\frac{l}{2}\right)}{\sqrt{\pi}\,\Gamma\left(\frac{l-1}{2}\right)}.$$

Thus, the density of u is

$$p(u) = \frac{\Gamma\left(\frac{l}{2}\right)}{\sqrt{\pi}\,\Gamma\left(\frac{l-1}{2}\right)} (1 - u^2)^{\frac{l-3}{2}}, \quad u \in [-1, 1].$$

This completes the proof. □

Theorem 4. *The closed form of $MMD_N^2(\mathcal{P}, \mathcal{Q})$, where $\mathcal{P}$ is a uniform distribution on $\mathcal{S}^{l-1}$ can be computed analytically (with N samples) to yield the following:*

$$\textit{MMD}_N^2(\mathcal{P}_N, \mathcal{Q}_N) = -a(\sigma, l) + \frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{j \neq i}^{N} e^{-\frac{\|\mathbf{z}_i - \mathbf{z}_j\|^2}{2\sigma^2}} \tag{12}$$

and, our proposed regularizer is related with $\textit{MMD}_N^2$ as follows:

$$\log \textit{MMD}_N^2(\mathcal{P}_N, \mathcal{Q}_N) \leq \mathcal{L}_{DIME}^{info}.$$

Here, $a(\sigma, l) = e^{-\sigma^{-2}} \Gamma(l/2) \left(2\sigma^2\right)^{\frac{l-2}{2}} \mathbf{I}_{\frac{l-2}{2}}\left(\sigma^{-2}\right)$.

Proof. The population MMD can be most straight-forwardly computed via the formula (Gretton et al., 2012):

$$MMD^2(\mathcal{P}, \mathcal{Q}) = \mathbb{E}_{\mathbf{w}, \mathbf{w}' \sim \mathcal{P}}[k(\mathbf{w}, \mathbf{w}')] - 2\mathbb{E}_{\mathbf{w} \sim \mathcal{P}, \mathbf{z} \sim \mathcal{Q}}[k(\mathbf{w}, \mathbf{z})] + \mathbb{E}_{\mathbf{z}, \mathbf{z}' \sim \mathcal{Q}}[k(\mathbf{z}, \mathbf{z}')] \tag{13}$$

The target distribution $\mathcal{P}$ is a uniform distribution over sphere $\mathcal{S}^{l-1}$. Now, given the samples from $\mathcal{Q}$, the goal is to derive a closed form estimate of $MMD^2(\mathcal{P}, \mathcal{Q})$.

Unbiased Estimator: We start with the expression Eq.(13) and using the sample $\mathcal{Q}_N = \{\mathbf{z}_i\}_{i=1}^N$ of size N, we replace the last two terms by the sample average and the U-statistic respectively to obtain the unbiased estimator (Gretton et al., 2012):

$$MMD_N^2(\mathcal{P}, \mathcal{Q}) = \mathbb{E}_{\mathbf{w}, \mathbf{w}' \sim \mathcal{P}}[k(\mathbf{w}, \mathbf{w}')] - \frac{2}{N} \sum_{i=1}^{N} \mathbb{E}_{\mathbf{w} \sim \mathcal{P}}[k(\mathbf{w}, \mathbf{z}_i)] + \frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{j \neq i}^{N} k(\mathbf{z}_i, \mathbf{z}_j). \tag{14}$$

From Eq.(14), we will now show that the first two expectations can be computed in closed form. Let us start with the first term and rewrite it as an integral:

$$\begin{aligned} \mathbb{E}_{\mathbf{w}, \mathbf{w}' \sim \mathcal{P}}[k(\mathbf{w}, \mathbf{w}')] = &\quad \mathbb{E}_{\mathbf{w}, \mathbf{w}' \sim \mathcal{P}}\left[e^{-\frac{1}{2\sigma^2}\|\mathbf{w} - \mathbf{w}'\|}\right] \\ = &\quad e^{-\frac{1}{\sigma^2}} \mathbb{E}_{\mathbf{w}, \mathbf{w}' \sim \mathcal{P}}\left[e^{\mathbf{w}^\intercal \mathbf{w}'/\sigma^2}\right] \quad (\|\mathbf{w} - \mathbf{w}'\| = 2 - 2\mathbf{w}^\intercal \mathbf{w}') \end{aligned}$$

When $\mathbf{w}$ and $\mathbf{w}'$ are independent and uniformly distributed on $\mathcal{S}^{l-1}$, the inner product

$$u = \mathbf{w}^\intercal \mathbf{w}', \quad u \in [-1, 1],$$

has the probability density

$$p(u) = \frac{\Gamma(l/2)}{\sqrt{\pi}\Gamma\left(\frac{l-1}{2}\right)}(1-u^2)^{\frac{l-3}{2}}, \quad u \in [-1,1].$$

This result is derived from **Lemma** 1. Thus, we can write

$$\mathbb{E}_{\mathbf{w},\mathbf{w}'\sim\mathcal{P}}\left[e^{\mathbf{w}^\intercal\mathbf{w}'/\sigma^2}\right] = \int_{-1}^{1} e^{u/\sigma^2}p(u)du$$

A standard integral formula (related to the Funk–Hecke theorem) states that for $\alpha > 0$ and for parameters such that the integral converges,

$$\int_{-1}^{1} e^{\alpha u}(1-u^2)^{\frac{l-3}{2}}du = \sqrt{\pi}\Gamma\left(\frac{l-1}{2}\right)(2/\alpha)^{\frac{l-2}{2}}\mathbf{I}_{\frac{l-2}{2}}(\alpha),$$

where $\mathbf{I}_\nu(.)$ is the modified Bessel function of the first kind. Here, we set

$$\alpha = \frac{1}{\sigma^2}.$$

Then the integral becomes

$$\int_{-1}^{1} e^{u/\sigma^2}(1-u^2)^{\frac{l-3}{2}}du = \sqrt{\pi}\Gamma\left(\frac{l-1}{2}\right)(2\sigma^2)^{\frac{l-2}{2}}\mathbf{I}_{\frac{l-2}{2}}(1/\sigma^2).$$

Multiplying by the constant in the density $p(u)$, we have

$$\begin{aligned}\mathbb{E}_{\mathbf{w},\mathbf{w}'\sim\mathcal{P}}\left[e^{\mathbf{w}^\intercal\mathbf{w}'/\sigma^2}\right] &= \frac{\Gamma(l/2)}{\sqrt{\pi}\Gamma\left(\frac{l-1}{2}\right)}\left[\sqrt{\pi}\Gamma\left(\frac{l-1}{2}\right)(2\sigma^2)^{\frac{l-2}{2}}\mathbf{I}_{\frac{l-2}{2}}(1/\sigma^2)\right] \\ &= \Gamma(l/2)(2\sigma^2)^{\frac{l-2}{2}}\mathbf{I}_{\frac{l-2}{2}}(1/\sigma^2).\end{aligned}$$

Finally, the first expectation can be written as

$$\mathbb{E}_{\mathbf{w},\mathbf{w}'\sim\mathcal{P}}[k(\mathbf{w},\mathbf{w}')] = e^{-\frac{1}{\sigma^2}}\Gamma(l/2)(2\sigma^2)^{\frac{l-2}{2}}\mathbf{I}_{\frac{l-2}{2}}(1/\sigma^2)$$

Now, interestingly, because of rotation–invariance of the spherical uniform distribution, the distribution of the inner product $u_i = \mathbf{w}^\intercal\mathbf{z}_i$ is unchanged from the previous case—even if $\mathbf{z}_i$ is fixed and $\mathbf{w}$ is uniformly distributed over $\mathcal{S}^{l-1}$. Thus, the same calculation applies and we obtain the same closed-form formulae for the second expectation also:

$$\mathbb{E}_{\mathbf{w}\sim\mathcal{P}}[k(\mathbf{w},\mathbf{z}_i)] = e^{-\frac{1}{\sigma^2}}\Gamma(l/2)(2\sigma^2)^{\frac{l-2}{2}}\mathbf{I}_{\frac{l-2}{2}}(1/\sigma^2) \quad \forall i \in [N]$$

Hence, the final closed form formulae for $MMD_N^2(\mathcal{P},\mathcal{Q})$, where $\mathcal{P}$ is a uniform distribution over $l-1$ dimensional hypersphere is:

$$MMD_N^2(\mathcal{P},\mathcal{Q}) = -\underbrace{e^{-\frac{1}{\sigma^2}}\Gamma(l/2)(2\sigma^2)^{\frac{l-2}{2}}\mathbf{I}_{\frac{l-2}{2}}(1/\sigma^2)}_{a(\sigma,l)} + \underbrace{\frac{1}{N(N-1)}\sum_{i=1}^{N}\sum_{j\neq i}^{N}k(\mathbf{z}_i,\mathbf{z}_j)}_{e^{\mathcal{L}_{DIME}^{info}}} \tag{15}$$

As, $l, d > 0$, implies $e^{-\frac{1}{\sigma^2}}\Gamma(l/2)(2\sigma^2)^{\frac{l-2}{2}}\mathbf{I}_{\frac{l-2}{2}}(1/\sigma^2) > 0$. Hence, $\mathcal{L}_{DIME}^{info} \geq \log MMD_N^2(\mathcal{P},\mathcal{Q})$. This completes the proof. □

A.5 DRAWBACKS OF INFOVAE

Info-VAEs proposed to optimize the following objective

$$\max_{\phi,\theta}\mathbb{E}_{q(\mathbf{x})}\Big[\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\big[\log p_\theta(\mathbf{x}|\mathbf{z})\big]\Big] + (\alpha-\beta)\,\mathrm{KL}(q_\phi(\mathbf{z}|\mathbf{x})\|p(\mathbf{z})) - \alpha\,\mathrm{KL}(q_\phi(\mathbf{z})\|p(\mathbf{z})).$$

to circumvent calculating the mutual information. This paper proposes an approach that directly optimizes a KDE-based mutual information using a VMF kernel. Specifically, DIME estimates mutual information by incorporating an explicit regularizer to maximize the dependency between the data and the latent space of a simple and deterministic autoencoder. This ensures a strong relationship between the input and latent codes, enhancing the quality of learned representations, which is clearly observed in **Table** 1 and **Table** 2 of our main manuscript.

More important than the objectives, InfoVAE is limited to $\alpha \leq 1$ since they circumvent the mutual information; otherwise, the term $\mathrm{KL}(q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z}))$ blows up in the first iteration since the encoder immediately learns with $\sigma_{\mathbf{z}|\mathbf{x}} = 0$. The InfoVAE's objective becomes infinity (as $\mathrm{KL}(q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z})) \to \infty$). Note that the InfoVAE's objective is:

$$\max_{\phi,\lambda}\mathbb{E}_{q(\mathbf{x})}\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\big[\log p_\theta(\mathbf{x}|\mathbf{z})\big] - (1-\alpha)\mathrm{KL}(q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z})) - (\alpha+\lambda)\mathrm{KL}(q_\phi(\mathbf{z})||p(\mathbf{z})).$$

Therefore, the amount of useful information in latent variables for InfoVAE is limited because of the restrictions imposed by $(1-\alpha)\mathrm{KL}(q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z}))$. However, our proposed InfoMax does not require $\alpha \leq 1$. Note that α plays a critical role as it determines the information preference in VAEs.

A.6 More Qualitative Results

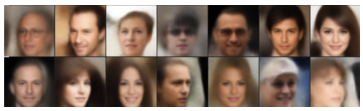
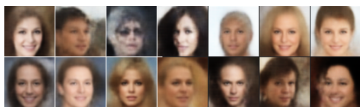
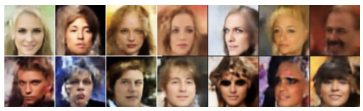
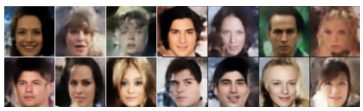
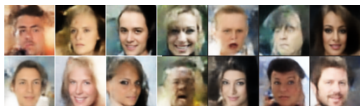
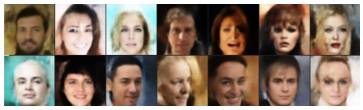
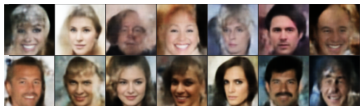
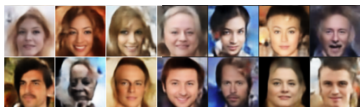
Model	Distribution	Generated Images
VAE	$\mathcal{N}(\mathbf{0}, \mathbb{I})$	
INFO-VAE	$\mathcal{N}(\mathbf{0}, \mathbb{I})$	
AE	GMM	
	MVG	
WAE	GMM	
	$\mathcal{N}(\mathbf{0}, \mathbb{I})$	
RAE-L_2	GMM	
	MVG	
RAE-GP	GMM	
	MVG	
DIME	GMM	
	MVG	

Table 8: A qualitative comparison of various autoencoder models for image generation on the CelebA dataset, highlighting their performance and characteristics.

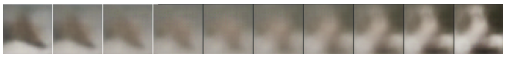
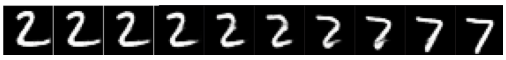
Model	Dataset	Interpolation
VAE	MNIST	
	CIFAR-10	
	CelebA	
InfoVAE	MNIST	
	CIFAR-10	
	CelebA	
AE	MNIST	
	CIFAR-10	
	CelebA	
WAE	MNIST	
	CIFAR-10	
	CelebA	
RAE-L_2	MNIST	
	CIFAR-10	
	CelebA	
RAE-GP	MNIST	
	CIFAR-10	
	CelebA	
DIME	MNIST	
	CIFAR-10	
	CelebA	

Table 9: Interpolation results for different models across datasets

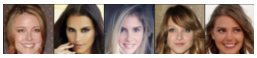
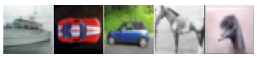
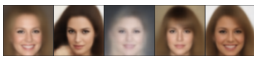
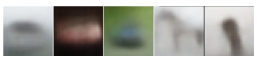
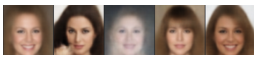
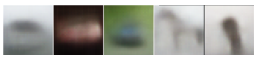
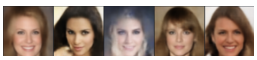
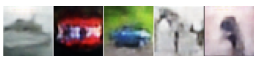
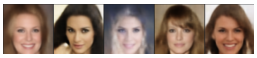
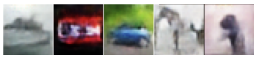
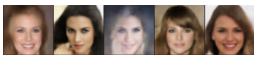
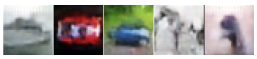
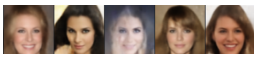
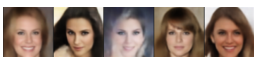
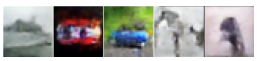
Model	CelebA	MNIST	CIFAR10
Original			
VAE			
INFO-VAE			
AE			
WAE			
RAE-L_2			
RAE-GP			
DIME			

Table 10: A qualitative comparison of the reconstruction quality achieved by different autoencoder models across the CelebA, MNIST, and CIFAR10 datasets.

A.7 Effect of tuning hyperparameters on FID score

In this section, we investigate three crucial hyperparameters: **(i)** penalty parameter (λ), **(ii)** encoder output dimension (latent dimension), and **(iii)** the precision parameter of Gaussian kernel (t). We seek insights into their impact on DIME's generative capacity. All experiments are conducted on the CelebA dataset, where two parameters are fixed while the third is varied. The fixed parameters are set to their default values as specified in **Table** 7.

A.7.1 Effect of penalty parameter λ on FID score

Elevating the penalty term results in a more pronounced regularization effect, whereas reducing it lessens this impact. We investigate the effect of the regularization penalty parameter on FID scores.

λ	1	0.5	0.1	5×10^{-2}	10^{-2}	5×10^{-3}	10^{-3}	5×10^{-4}
MVG	47.08	43.7	46.93	49.55	45.30	47.79	54.72	55.47
GMM	41.84	38.48	41.54	44.57	39.36	42.75	48.66	49.89

Table 11: Generative performance of DIME on CelebA dataset across different values of λ, where $t = 5 \times 10^{-3}$ and $l = 64$.

A.7.2 Effect of latent dimension on FID score

Latent Dimension (l)	16	32	64	128	256
MVG	70.55	53.03	45.30	50.19	53.52
GMM	65.23	48.16	39.36	44.73	47.08

Table 12: Generative performance of DIME on CelebA dataset across different values of l, where $t = 5 \times 10^{-3}$ and $\lambda = 10^{-2}$.

A.7.3 Effect of precision parameter (t) on FID score

t	0.1	5×10^{-2}	10^{-2}	5×10^{-3}	10^{-3}	5×10^{-4}	10^{-4}
MVG	46.89	49.07	49.38	45.30	46.58	46.92	48.54
GMM	41.67	46.55	44.78	39.36	41.11	41.70	43.01

Table 13: Generative performance of DIME on CelebA dataset across different values of t, where $\lambda = 10^{-2}$ and $l = 64$.