Intervention Analysis on the Latent Space of Variational Autoencoders

Advanced Machine Learning Project

Alexander Müller, (s5193400)
Adrian Predescu, (s4729307)
Tim Ludwig, (s4665198)
Mika Umaña Lemus, (s5173213)
Aalam Sultanji, (s5130468)

Abstract

This paper investigates how the diversity of training data influences uncertainty and reconstruction blur in Variational Autoencoders (VAEs). Using controlled subsets of MNIST and FashionMNIST, we vary the number of classes and samples per class to create datasets with different diversity levels, measured by the Structural Similarity Index (SSIM). We then train identical VAEs and apply an intervention analysis to identify which latent dimensions contribute most to reconstruction error. Our results show that higher dataset diversity and reconstruction ratio sum leads to a richer latent space, where uncertainty spreads across more distinct dimensions. Moreover, interventions work better on models trained with diverse data, showing that variety in the training set makes the VAE's latent space more interpretable and responsive to change. These findings indicate that having a diverse training set causes more significant dimensions in the latent space of VAEs with respect to reconstruction error. Code is made available at https://github.com/AlexanderAKM/Advanced_ML.

Keywords

variational autoencoder; epistemic uncertainty; latent space; reconstruction sharpness; counterfactual analysis; dataset diversity

1 Introduction

1.1 Introduction to VAE

Variational Autoencoders are widely used tools [1], as they are very efficient in detecting the underlying data distribution and are able to generate new, realistic samples. Their architecture consists of a probabilistic approach and an encoder-decoder framework, which makes them highly versatile and adaptable to various tasks.

The use cases of VAEs span a wide range of domains, particularly in generating and manipulating media. For example, a VAE trained on a dataset of faces can create new, unique human faces that look realistic [2]. This capability is valuable in creative arts, entertainment, advertising, and virtual reality. Beyond generative tasks, VAEs are also effective for tasks like image denoising, enhancement, and even generating 3D data like point clouds and meshes [3][4].

The usage of VAEs in this context is very efficient, as its latent space is a compressed, lower-dimensional representation of the input data [5]. Unlike a standard autoencoder that creates a fixed encoding, a VAE's encoder maps the input to a probability distribution within this latent space [4]. By sampling from this continuous latent distribution and passing the sample to the decoder, VAEs can generate new data instances that resemble the original training data [2]. This probabilistic approach allows for the generation of diverse and novel samples rather than just reconstructing inputs [2].

Further use cases of VAEs that are relevant,

but not applicable to this project include but are not limited to: anomaly detection, molecular generation, and financial modeling.

In anomaly detection, VAEs are trained on "normal" data to learn typical patterns and later identify deviations, which is useful for fraud detection in finance or defect detection in manufacturing [6][7]. In molecular design, VAEs trained on large chemical datasets can generate new molecular structures with desired properties, helping drug discovery [8]. In finance, VAEs have been applied to model and generate synthetic volatility surfaces for derivative pricing and hedging [9], or to synthesize realistic financial transaction data when real-world examples, such as fraud cases, are limited [10].

Finally, VAEs are valuable for data augmentation, i.e. artificially expanding datasets with realistic synthetic samples to improve model generalization [1]. They are also used for dimensionality reduction, compressing high-dimensional data into compact latent representations while preserving the key structural information [9].

1.2 Architecture of a VAE

A Variational Autoencoder is a probabilistic generative model that leverages a neural network-based encoder-decoder architecture [11][12]. Its primary objective is to learn a probabilistic mapping from a high-dimensional data space $\mathcal{X}$ to a lower-dimensional, continuous latent space $\mathcal{Z}$, from which new data can be generated.

1.2.1 The Generative Model (Decoder)

The generative process, or decoder, defines a distribution over the data $\mathbf{x} \in \mathcal{X}$ conditioned on a latent variable $\mathbf{z} \in \mathcal{Z}$. This process is defined by two components:

1. A prior distribution over the latent variables, $p(\mathbf{z})$, which is typically chosen to be a standard multivariate Gaussian: $p(\mathbf{z}) = \mathcal{N}(\mathbf{z}; \mathbf{0}, \mathbf{I})$.
2. A conditional distribution $p_\theta(\mathbf{x}|\mathbf{z})$, representing the likelihood of observing $\mathbf{x}$ given $\mathbf{z}$. This distribution is parameterized by a neural network (the decoder) with parameters θ. For real-valued data like images, $p_\theta(\mathbf{x}|\mathbf{z})$ is often modeled as a Gaussian $\mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_\theta(\mathbf{z}), \text{diag}(\boldsymbol{\sigma}^2_\theta(\mathbf{z})))$, or a Bernoulli distribution for binary data.

The ultimate goal is to maximize the marginal likelihood of the data, $p_\theta(\mathbf{x}) = \int p_\theta(\mathbf{x}|\mathbf{z})p(\mathbf{z})d\mathbf{z}$. However, this integral is generally intractable due to its high dimensionality and the complexity of $p_\theta(\mathbf{x}|\mathbf{z})$.

1.2.2 The Inference Model (Encoder)

To overcome the intractability of the true posterior $p_\theta(\mathbf{z}|\mathbf{x}) = \frac{p_\theta(\mathbf{x}|\mathbf{z})p(\mathbf{z})}{p_\theta(\mathbf{x})}$, VAEs introduce a variational approximation, $q_\phi(\mathbf{z}|\mathbf{x})$, parameterized by another neural network (the encoder) with parameters ϕ [13]. This approximate posterior is designed to be a tractable distribution, typically a multivariate Gaussian with a diagonal covariance structure:

$$q_\phi(\mathbf{z}|\mathbf{x}) = \mathcal{N}(\mathbf{z}; \boldsymbol{\mu}_\phi(\mathbf{x}), \text{diag}(\boldsymbol{\sigma}^2_\phi(\mathbf{x}))) \tag{1.1}$$

Here, the encoder network takes an input $\mathbf{x}$ and outputs the mean vector $\boldsymbol{\mu}_\phi(\mathbf{x})$ and the log-variance vector $\log \boldsymbol{\sigma}^2_\phi(\mathbf{x})$.

1.2.3 The Objective Function: Evidence Lower Bound (ELBO)

The parameters ϕ and θ are jointly optimized by maximizing the Evidence Lower Bound (ELBO), denoted $\mathcal{L}(\phi, \theta; \mathbf{x})$, on the marginal log-likelihood [11]. The ELBO is derived as follows:

$$\log p_\theta(\mathbf{x}) = \log \int p_\theta(\mathbf{x}, \mathbf{z}) d\mathbf{z} \tag{1.2}$$

$$= \log \int p_\theta(\mathbf{x}, \mathbf{z}) \frac{q_\phi(\mathbf{z}|\mathbf{x})}{q_\phi(\mathbf{z}|\mathbf{x})} d\mathbf{z} \tag{1.3}$$

$$= \log \mathbb{E}_{\mathbf{z} \sim q_\phi(\mathbf{z}|\mathbf{x})} \left[\frac{p_\theta(\mathbf{x}, \mathbf{z})}{q_\phi(\mathbf{z}|\mathbf{x})} \right] \tag{1.4}$$

$$\geq \mathbb{E}_{\mathbf{z} \sim q_\phi(\mathbf{z}|\mathbf{x})} \left[\log \frac{p_\theta(\mathbf{x}, \mathbf{z})}{q_\phi(\mathbf{z}|\mathbf{x})} \right] \quad \text{(by Jensen's Inequality)} \tag{1.5}$$

This lower bound can be rewritten in a more intuitive form [14]:

$$\mathcal{L}(\phi, \theta; \mathbf{x}) = \underbrace{\mathbb{E}_{\mathbf{z}\sim q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})]}_{\text{Reconstruction Term}} - \underbrace{D_{KL}(q_\phi(\mathbf{z}|\mathbf{x})\|p(\mathbf{z}))}_{\text{Regularization Term}} \tag{1.6}$$

The first term is the expected negative reconstruction error. It forces the decoder to learn to reconstruct the input data $\mathbf{x}$ from its latent representation $\mathbf{z}$. The second term is the Kullback-Leibler (KL) divergence between the approximate posterior and the prior. This term acts as a regularizer, encouraging the encoder to learn distributions $q_\phi(\mathbf{z}|\mathbf{x})$ that are close to the standard normal prior, which enforces a smooth and well-structured latent space. For a Gaussian encoder and prior, this KL divergence term has a convenient closed-form analytical solution:

$$D_{KL}(\mathcal{N}(\boldsymbol{\mu}, \text{diag}(\boldsymbol{\sigma}^2)) \,\|\, \mathcal{N}(\mathbf{0}, \mathbf{I})) = \frac{1}{2}\sum_{j=1}^{J}\left(\sigma_j^2 + \mu_j^2 - 1 - \log\sigma_j^2\right) \tag{1.7}$$

where J is the dimensionality of the latent space.

1.2.4 Optimization via the Reparameterization Trick

A challenge in optimizing the ELBO is that the gradient cannot be backpropagated through the sampling step $\mathbf{z} \sim q_\phi(\mathbf{z}|\mathbf{x})$. The *reparameterization trick* resolves this issue by reframing the sampling process [11]. Instead of sampling $\mathbf{z}$ directly, we sample a noise vector $\boldsymbol{\epsilon}$ from a simple distribution, such as $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, and then compute $\mathbf{z}$ as a deterministic function of $\boldsymbol{\mu}_\phi$, $\boldsymbol{\sigma}_\phi$, and $\boldsymbol{\epsilon}$:

$$\mathbf{z} = \boldsymbol{\mu}_\phi(\mathbf{x}) + \boldsymbol{\sigma}_\phi(\mathbf{x}) \odot \boldsymbol{\epsilon} \tag{1.8}$$

where $\odot$ denotes the element-wise product. This formulation externalizes the stochasticity, allowing gradients to flow from the objective function back through $\mathbf{z}$ to the parameters ϕ and θ of the encoder and decoder, respectively. The final loss function to be minimized via stochastic gradient descent is the negative of the ELBO, where the expectation is approximated using a single Monte Carlo sample:

$$L(\phi, \theta; \mathbf{x}) \approx -\log p_\theta(\mathbf{x}|\mathbf{z}) + D_{KL}(q_\phi(\mathbf{z}|\mathbf{x})\|p(\mathbf{z})) \tag{1.9}$$

where $\mathbf{z}$ is obtained using the reparameterization trick

1.3 Uncertainty in VAEs and its Consequences

With all the benefits VAEs bring to the table, they also have several inherent limitations, most notably uncertainty, which can degrade their performance. This uncertainty often manifests as blurry image generation and can contribute to posterior collapse, where the latent variables fail to capture the meaningful information about the input data [15][16].

At its core, a VAE is a generative model that learns the probabilistic mapping between the input data and a lower-dimensional latent space. Unlike a standard autoencoder that learns a single, fixed encoding for each input, a VAE's encoder outputs a probability distribution (typically a Gaussian) for each input. This distribution is defined by a mean and a variance, where the variance represents the uncertainty about the latent representation of a given data point. The decoder then samples from this distribution to generate new data. This probabilistic nature is what allows VAEs to generate diverse outputs [17], but also the reason why image sharpness can be hindered.

1.4 Variety of Image Generation in VAEs

As mentioned, a key strength of VAEs is their ability to generate diverse, realistic images by using a structured latent space. Unlike standard autoencoders that produce deterministic reconstructions, VAEs introduce controlled randomness through probabilistic encoding, being able to create novel outputs beyond the training data [3]. This generative diversity is facilitated by the following core concepts:

- **Latent Space Sampling:** The latent space in a VAE is a low-dimensional, continuous representation of the input data [3]. Because the encoder outputs a distribution rather than a single point, it introduces a degree of controlled randomness [18]. This means that for a single input image, the encoder describes a region in the latent space from which a latent vector can be sampled [19][20].

- **Sampling for Novelty:** To generate a new image, a random point is sampled from this learned distribution in the latent space and then passed to the decoder [20]. The decoder, which is a neural network trained to reconstruct images from these latent vectors, then generates an image that is similar, but not identical, to the original training data [20]. This sampling process is what allows VAEs to produce a continuous spectrum of new outputs [19].

- **KL Regularization:** During training, a regularization term, known as the Kullback-Leibler (KL) divergence, encourages the learned distributions in the latent space to be close to a standard normal distribution [21][22]. This regularization helps to create a smooth and well-structured latent space where nearby points correspond to visually similar images [3]. This continuity allows for meaningful interpolation; by selecting two points in the latent space and decoding the intermediate points, one can generate a smooth transition between the corresponding images [22].

In essence, the VAE learns the underlying probability distribution of the training data [21]. By sampling from this learned distribution, the VAE can generate a multitude of new images that exhibit the same characteristics as the training set, effectively "dreaming up" new variations of the data it has seen [19]. This ability to generate diverse and plausible outputs makes VAEs a powerful tool in various creative AI applications [3].

1.5 Uncertainty vs. Variety Tradeoff

A fundamental challenge in working with Variational Autoencoders lies in balancing the "uncertainty" in the latent space with the "variety" or quality of the generated samples. This tradeoff is primarily governed by the two main components of the VAE's loss function: the reconstruction loss and the Kullback-Leibler (KL) divergence [23][24][25][26][27].

The reconstruction loss encourages the VAE to accurately reconstruct the input data from its latent representation [23][25][26]. Minimizing this loss leads to higher-fidelity reconstructions, meaning the generated samples are sharp and closely resemble the original data. However, focusing solely on reconstruction can lead to a less structured and more "certain" latent space, where the model essentially learns to memorize the training data, limiting its ability to generate novel and diverse samples.

On the other hand, the KL divergence term acts as a regularizer, pushing the distribution of the learned latent variables to be close to a predefined prior distribution, typically a standard normal distribution [23][25]. This regularization encourages a more structured and "uncertain" latent space, preventing overfitting and enabling the generation of a wider variety of new data points [23]. However, placing too much emphasis on the KL divergence can come at the cost of reconstruction quality, leading to blurry and less detailed generated images [23][25].

This inherent tension creates a tradeoff: a model that excels at reconstruction may lack generative diversity, while a model that prioritizes a smooth and regularized latent space might produce less realistic outputs [26]. The β-VAE directly addresses this tradeoff by introducing a weighting factor (β) to the KL divergence term, allowing researchers to explicitly control this balance [28][29]. Finding the optimal balance between these two competing objectives is a key aspect of training effective VAEs and is often dependent on the specific application and desired outcome.

1.6 Variety of Datasets

Datasets can have different varieties and complexities of data, from very simple datasets, e.g. Iris Petal Dataset [30] with only very few datapoints and only five features, to very complex datasets that include almost all of the internet and are used to train Large Language Models (LLMs), or Gene Expression Datasets with over 20.000 features like the Gene Expression Omnibus dataset [31] which are know for the high complexity.

Given this diverse landscape, it is paramount to recognize that not all AI models are suited for all types of data. A very complex model might not be able to learn a simple representation like in the Iris Petal dataset, if only very few data points are available, but will definitely be an overkill and waste resources and computational power. The final performance of the task that one wants to perform is highly dependent on the choice of model [32].

Datasets might also be very homogeneous, for example repetition of data points, large over-representation of a certain class, similar images (in an image dataset) etc. This homogeneity can lead to issues like overfitting, reduced model effectiveness and increased bias [33].

VAEs specifically have a tendency to struggle with homogeneous datasets. When the training data lacks diversity, the VAE may learn a latent space that is not very rich or expressive, potentially leading to overfitting where the model essentially memorizes the limited variations in the data [34]. This restricts its ability to generate novel and diverse samples, which is one of its key strengths. For instance, if a VAE is trained only on very similar images, the learned probability distributions in the latent space will be tightly clustered. Interventions aimed at reducing reconstruction blurriness by averaging across certain latent dimensions might have a limited effect, as there is little variety to begin with. On the other hand, in more complex and diverse datasets, the VAE learns a more structured latent space where different dimensions capture more distinct features. In this scenario, intervening in the latent space to reduce uncertainty can significantly decrease reconstruction blur, but it may also diminish the output's variety, a crucial trade-off. This is why selecting datasets with sufficient diversity is critical for analyzing the relationship between latent uncertainty and reconstruction quality.

To quantitatively check the variety of a dataset, we employ the Structural Similarity Index (SSIM) as our diversity metric, see 2.2. Rather than relying on pre-trained feature extractors like ResNet-18, which we ultimately have not used, see Appendix B, we directly measure dissimilarity between image pairs. Specifically, for each dataset, we sample images and compute the SSIM between all pairs, which capture different characteristics of the images. The diversity score is then defined as a dissimilarity measure, see Section 2.2. A higher score indicates greater visual and semantic heterogeneity, signifying a more diverse dataset with a wider range of features and content [35][36].

1.7 Research Goal

To summarize, VAEs are powerful generative models, though their complex, high-dimensional latent spaces can make them behave like "black boxes," making it difficult to interpret their internal representations and diagnose performance issues [37]. A widely acknowledged limitation, for instance, is their tendency to produce blurry reconstructions, a side effect of the distributional averaging inherent in their objective function [38]. The conventional approach to mitigate such issues—iterating on model architectures or engaging in extensive hyperparameter tuning—is often computationally expensive with no guarantee of success [39]. This challenge is being made more complex by the fact that selecting an optimal model architecture for a given dataset's complexity remains a significant open problem in machine learning.

This abundance of challenges highlights a critical need for methods that can analyze and improve model performance post-training. Such post-hoc analyses, which probe a model's learned representations after training is com-

plete, offer a more resource-efficient path toward understanding model behavior and enhancing reliability without the need for complete re-training [40]. This paper adopts such a post-training approach by investigating the fundamental relationship between the diversity of the training data and the manifestation of epistemic uncertainty in VAE reconstructions.

Our central goal is to understand how the structure of the learned latent space is shaped by data diversity and how this, in turn, influences reconstruction quality. To this end, we formulate a primary hypothesis:

- **Hypothesis**: The effectiveness of post-training interventions to reduce reconstruction blur is contingent on the diversity of the training data. For low-diversity (homogeneous) datasets, a VAE will learn a tightly clustered latent space where uncertainty is diffuse, and interventions will have a limited and uniform effect. Conversely, for high-diversity (complex) datasets, the VAE will learn a more structured latent space where distinct dimensions capture meaningful features. In this scenario, targeted interventions can significantly reduce uncertainty-driven blur, potentially at the cost of output variety.

To systematically test this hypothesis, we aim to:

- Propose and apply a quantitative framework for measuring dataset diversity. This is achieved by aggregating pairwise visual dissimilarities across the dataset, using the well-established Structural Similarity Index (SSIM) as the foundational metric.

- Analyze how VAEs trained on datasets with controlled levels of diversity represent uncertainty in their latent space.

- Employ a targeted intervention analysis to identify which latent dimensions are the primary contributors to reconstruction blur and examine how this contribution varies with dataset diversity.

By pursuing these objectives, we want to move beyond training models and instead develop a deeper understanding of how data properties influence latent representations.

This work aims to contribute to post-hoc techniques that can enhance VAE performance, saving valuable computational resources and providing clearer insights into the internal workings of these generative models.

2 Methodology

2.1 Experimental Design Overview

To systematically investigate the relationship between dataset characteristics and the manifestation of uncertainty in VAE reconstructions, we designed a multi-dataset experimental framework. Our approach consists of three key components:

1. Quantitative measurement of dataset diversity

2. Training of VAE models across multiple subsets of our datasets

3. Intervention-based analysis of latent uncertainty

To achieve this, we create a controlled-diversity framework in which the variation within the training data is explicitly determined by the number of classes (C) and the number of samples per class (S_c). We evaluate three diversity levels:

$$(C, S_c) \in \{(1, 1000),\ (5, 1000),\ (10, 1000)\}$$

All models use the same architecture, optimizer, and hyperparameters. Additionally, to make sure that any observed variation arises purely from dataset composition, we repeat each configuration for $N = 10$ independent runs. For each run, we vary the dataset seed $s_d \in [2810, 2819]$ to create different random subsets, while keeping the training seed fixed at $s_t = 2810$. This approach makes it easy to

control the influence of random sampling while maintaining identical model initialization and learning dynamics across runs (more information on the hyperparameters can be found in Appendix A).

In an initial attempt, we tried to quantify dataset diversity using a feature-based metric derived from ResNet-18 embeddings on nine grayscale datasets. While the idea seemed feasible, we found that it did not reliably capture meaningful differences for small, low-resolution, grayscale images, giving redundant results (more details in Appendix B).

2.2 Dataset Selection and Diversity Measurement

We selected two grayscale image datasets to ensure sufficient variability in visual complexity and content diversity:

- **MNIST** [41]: A cornerstone of machine learning, the MNIST (Modified National Institute of Standards and Technology) dataset is a collection of 70,000 handwritten digits, from 0 to 9. It is divided into 60,000 images for training and 10,000 for testing. Each image is a 28x28 pixel grayscale representation. This dataset was created by "re-mixing" samples from NIST's original datasets to better suit machine learning experiments. It has become a fundamental benchmark for image processing and classification algorithms.
- **FashionMNIST** [42]: Created as a more challenging drop-in replacement for the original MNIST dataset, FashionMNIST features 70,000 grayscale images of fashion products from 10 different categories. The dataset, provided by Zalando, is also split into 60,000 training and 10,000 testing images, each at a 28x28 pixel resolution. It serves as a benchmark for modern machine learning algorithms, offering a greater challenge than the classic handwritten digits.

For each dataset, we construct subsets by randomly selecting C classes and S_c examples per class with a fixed random seed for reproducibility. Each resulting subset therefore contains

$$N_{\text{total}} = C \times S_c$$

training samples. Moreover, all further diversity and intervention analyses are performed on the training subsets, as the goal is to examine how dataset composition influences latent-space structure rather than model generalization (throughout the rest of the paper, whenever we refer to a "dataset", we mean one of these controlled training subsets).

We then quantify dataset diversity using the Structural Similarity Index Measure (SSIM) [43], which measures perceptual similarity between pairs of images. For two normalized grayscale images $x_i, x_j \in [0,1]^{28\times 28}$, the SSIM is computed as

$$\text{SSIM}(x_i, x_j) = \frac{(2\mu_i\mu_j + c_1)(2\sigma_{ij} + c_2)}{(\mu_i^2 + \mu_j^2 + c_1)(\sigma_i^2 + \sigma_j^2 + c_2)},$$

where μ_i and μ_j denote the mean intensities, σ_i^2 and σ_j^2 are the sample variances, σ_{ij} is the sample covariance between x_i and x_j, $c_1 = (k_1 L)^2, c_2 = (k_2 L)^2$ are two variables to stabilize the division where L is the dynamic range of the pixel-values (typically $2^{\#\text{bits per pixel}} - 1$), and $k_1 = 0.01$ and $k_2 = 0.03$ by default [43]. To express structural dissimilarity, we define

$$d(x_i, x_j) = 1 - \text{SSIM}(x_i, x_j).$$

We then estimate the overall diversity of a dataset as the average dissimilarity across all pairwise combinations of a randomly sampled subset of images (200 samples per dataset for computational efficiency since SSIM is quite expensive):

$$D = \frac{2}{n(n-1)} \sum_{i<j} d(x_i, x_j),$$

and compute its sample variance as

$$\text{Var}(D) = \text{Var}_{i,j}\left[\, d(x_i, x_j) \,\right].$$

A higher value of D indicates a more diverse dataset in terms of content and visual variability.

2.3 Model Architecture and Training

For each of the datasets, we train an identical VAE architecture to ensure controlled comparison. The model consists of an encoder, reparameterization, and a decoder.

The encoder is a fully-connected feedforward network that maps the flattened 28×28 input image (784 dimensions) through a hidden layer of 400 units with ReLU activation, outputting the parameters of the approximate posterior distribution: mean $\boldsymbol{\mu} \in \mathbb{R}^{20}$ and log-variance $\log \boldsymbol{\sigma}^2 \in \mathbb{R}^{20}$.

Latent samples are drawn using the reparameterization trick [11]:

$$\mathbf{z} = \boldsymbol{\mu} + \boldsymbol{\sigma} \odot \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$$

The decoder is a feedforward network mapping the 20-dimensional latent code through a 400-unit hidden layer with ReLU activation, followed by a final layer of 784 units with sigmoid activation to reconstruct the image.

The model is trained by maximizing the evidence lower bound (ELBO):

$$\mathcal{L}(\mathbf{x}) = \mathbb{E}_{q(\mathbf{z}|\mathbf{x})}[\log p(\mathbf{x}|\mathbf{z})] - \beta \cdot D_{\mathrm{KL}}(q(\mathbf{z}|\mathbf{x}) \| p(\mathbf{z}))$$

where the reconstruction term uses binary cross-entropy loss:

$$\log p(\mathbf{x}|\mathbf{z}) = -\mathrm{BCE}(\hat{\mathbf{x}}, \mathbf{x})$$

and the KL divergence regularization is:

$$D_{\mathrm{KL}}(q(\mathbf{z}|\mathbf{x}) \| p(\mathbf{z})) = -\frac{1}{2} \sum_{j=1}^{20} \left(1 + \log \sigma_j^2 \quad - \mu_j^2 - \sigma_j^2\right)$$

We employ the Adam optimizer [44] with learning rate 10^{-3}, batch size 256, and train for 10 epochs.We had ambitions of trying our experimental design on a convolutional VAE, of which the implementation can be found on our GitHub, and the details are shown in Appendix C. Unfortunately, due to complexity and time factors, we were not able to properly show the results in this report.

2.4 Intervention-Based Uncertainty Analysis

To investigate whether reconstruction blur originates from epistemic uncertainty in the latent space, we perform a systematic intervention analysis on each trained model. For each dataset, we analyze $M = 500$ randomly selected images.

2.4.1 Baseline: Stochastic Reconstruction

For each image $\mathbf{x}$, we encode it to obtain the posterior parameters $(\boldsymbol{\mu}, \boldsymbol{\sigma}^2)$. We then sample a latent code $\mathbf{z} \sim q(\mathbf{z}|\mathbf{x})$ and decode it to obtain a reconstruction $\hat{\mathbf{x}}_{\text{sample}}$. The baseline reconstruction error is:

$$\mathcal{E}_{\text{sample}} = \mathrm{MSE}(\hat{\mathbf{x}}_{\text{sample}}, \mathbf{x})$$

where MSE denotes mean squared error. This represents the typical reconstruction quality when sampling from the learned posterior.

2.4.2 Mean Ablation

To establish a reference for the effect of removing all stochastic sampling, we decode directly from the posterior mean:

$$\hat{\mathbf{x}}_{\text{mean}} = \mathrm{Decoder}(\boldsymbol{\mu})$$

The mean reconstruction error is:

$$\mathcal{E}_{\text{mean}} = \mathrm{MSE}(\hat{\mathbf{x}}_{\text{mean}}, \mathbf{x})$$

The difference $\Delta\mathcal{E}_{\text{base}} = \mathcal{E}_{\text{sample}} - \mathcal{E}_{\text{mean}}$ quantifies the total impact of latent uncertainty on reconstruction quality.

2.4.3 Per-Dimension Interventions

To isolate the contribution of uncertainty in each individual latent dimension, we perform targeted interventions. For each dimension $i \in \{1, \ldots, 20\}$:

1. Sample a latent vector $\mathbf{z} \sim q(\mathbf{z}|\mathbf{x})$
2. Intervene by setting $z_i \leftarrow \mu_i$ (replacing the sampled dimension with the mean)

3. Decode the modified latent vector: $\hat{\mathbf{x}}_{\text{int},i} = \text{Decoder}(\mathbf{z})$

4. Compute the intervention reconstruction error: $\mathcal{E}_{\text{int},i} = \text{MSE}(\hat{\mathbf{x}}_{\text{int},i}, \mathbf{x})$

The reduction in reconstruction error for dimension i is:

$$\Delta\mathcal{E}_i = \mathcal{E}_{\text{sample}} - \mathcal{E}_{\text{int},i}$$

We then calculate the *contribution ratio* for each dimension:

$$R_i = \frac{\Delta\mathcal{E}_i}{\Delta\mathcal{E}_{\text{base}}}$$

This ratio quantifies what fraction of the total uncertainty-induced error is attributable to dimension i. Dimensions with $R_i > \tau$ (where $\tau = 0.075$ is a threshold representing 1.5 times the uniform contribution $1/20 = 0.05$) are considered significant dimensions.

2.5 Cross-Dataset Analysis and Metrics

For each dataset, we aggregate the intervention results across all $M = 500$ images to compute dataset-level summary statistics:

- **Mean Sample Error** ($\overline{\mathcal{E}}_{\text{sample}}$): Mean reconstruction error with stochastic sampling.
- **Mean Posterior Error** ($\overline{\mathcal{E}}_{\text{mean}}$): Mean reconstruction error using posterior mean.
- **Reconstruction Difference** ($\Delta\overline{\mathcal{E}}$): $\overline{\mathcal{E}}_{\text{sample}} - \overline{\mathcal{E}}_{\text{mean}}$, measuring the overall impact of uncertainty.
- **Ratio Sum** ($\sum_{i=1}^{20} \overline{R}_i$): Sum of mean contribution ratios across all dimensions, indicating the degree to which individual dimensions additively explain uncertainty.
- **Significant Dimensions** (N_{sig}): Number of dimensions with $\overline{R}_i > 0.075$, identifying how many latent factors meaningfully contribute to uncertainty-driven blur.
- **Intervention Variance** ($\text{Var}(\overline{\mathcal{E}}_{\text{int},1}, \ldots, \overline{\mathcal{E}}_{\text{int},20})$): Variance in reconstruction errors across interventions, measuring heterogeneity in dimension importance.

Finally, we compute Pearson correlation coefficients between dataset diversity (from Section 2.2) and each of the intervention metrics to investigate whether datasets with greater visual diversity exhibit different patterns of latent uncertainty and reconstruction characteristics.

All results, including per-dataset metrics and correlation analyses, are stored in a structured CSV file for reproducibility and further statistical testing and potential future research.

2.6 Theoretical Framework

There have been various attempts to write a mathematical proof that is in line with the topic and research. One of these wrong attempts can be found in Appendix D. For the final proof we will first state some preliminaries which we will then conclude with the final theorem and proof

Definition 1 (Dataset Diversity Score)**.** *We construct the dataset diversity as mentioned in 2.2 using the Structural Similarity Index Measure (SSIM):*

$$SSIM(x_i, x_j) = \frac{(2\mu_i\mu_j + c_1)(2\sigma_{ij} + c_2)}{(\mu_i^2 + \mu_j^2 + c_1)(\sigma_i^2 + \sigma_j^2 + c_2)}, \tag{2.1}$$

with the same definitions of the variables as in the mentioned section. To express structural dissimilarity, we define:

$$d(x_i, x_j) = 1 - SSIM(x_i, x_j) \tag{2.2}$$

and the overall diversity of a dataset as previously done by:

$$D = \frac{2}{n(n-1)} \sum_{i<j} d(x_i, x_j). \tag{2.3}$$

Definition 2 (Reconstruction Error and Contribution Ratio)**.** *For a Variational Autoencoder with encoder $q(\mathbf{z}|\mathbf{x})$ producing posterior*

parameters $(\boldsymbol{\mu}, \boldsymbol{\sigma}^2) \in \mathbb{R}^{20} \times \mathbb{R}^{20}$*, decoder* $p(\mathbf{x}|\mathbf{z})$*, and data* $x \sim \mathcal{D}$ *we define:*

$$\bar{E}_{sample} = \mathbb{E}_{x\sim\mathcal{D}}\left[\mathbb{E}_{z\sim q(z|x)}\left[1 - SSIM(Dec(z), x)\right]\right] \tag{2.4}$$

$$\bar{E}_{mean} = \mathbb{E}_x\left\{1 - SSIM(Dec(\mu(x)), x)\right\} \tag{2.5}$$

$$\overline{\Delta E}_{base} = \bar{E}_{sample} - \bar{E}_{mean} \tag{2.6}$$

Furthermore, for each latent dimension $i \in \{1, \ldots, 20\}$*, define the intervention reconstruction error:*

$$\bar{E}_{int,i} = \mathbb{E}_{x\sim\mathcal{D}}\left[\mathbb{E}_{z\sim q(z|x)}\left[1 - SSIM(Dec(z^{i\leftarrow\mu_i}), x)\right]\right] \tag{2.7}$$

where $z^{i\leftarrow\mu}$ *denotes the latent vector* z *with dimension* i *replaced by* μ_i*.*

The contribution ratio is:

$$\bar{R}_i = \frac{\overline{\Delta E}_i}{\overline{\Delta E}_{base}} = \frac{\bar{E}_{sample} - \bar{E}_{int,i}}{\bar{E}_{sample} - \bar{E}_{mean}} \tag{2.8}$$

Hypothesis 3 (Empirical View on Dataset Diversity)**.** *As the dataset diversity* D *increases, at least* $\bar{R}_i$ *increases and thus equivalently* $\max_i \bar{R}_i$ *correlates positively with* D*.*

2.6.1 Main Result

Theorem 4 (Diversity Implies Significant Dimensions)**.** *Let* $\gamma > 0$ *be a diversity threshold and* $\tau > 0$ *be a significance threshold. If a dataset with Diversity Score* $D > \gamma$ *implies higher entropy* $H(X)$ *and the VAE preserves reconstruction quality via* $I(X; z) \propto H(X)$*, then there exists at least one latent dimension* $i^* \in \{1, \ldots, d\}$ *such that* $R_{i^*} > \tau$*, where:*

$$\tau < \frac{\alpha \cdot \gamma}{d} \tag{2.9}$$

and $\{1, \ldots, d\}$ *are the latent dimensions and* $\alpha \cdot \gamma > 0$*,* $\alpha > 0 \in \mathbb{R}$*.*

Proof. Assumption: Information Bottleneck: Let $I(X; z)$ denote the mutual information between data X and latent representation z. The encoder-decoder pair of a VAE obeys:

$$I(X; z) \leq H(X),$$

where $H(X)$ is the entropy f the dataset. When the dataset diversity D increases, $H(X)$ increases monotonically, so the encoder must increase the mutual information $I(X; z)$ to preserve the reconstruction quality. The increase is reflected by higher latent variance σ^2 and stronger per-dimension contributions R_{i^*}.

We proceed by contradiction.

Assume $D > \gamma > 0$. Since $H(X)$ grows with dataset diversity D, maintaining good reconstruction requires a proportional increase in mutual information $I(X; z)$. Because

$$I(X; z) = \mathbb{E}_x[D_{KL}(q(z|x)||p(z)]$$

a larger $I(X; z)$ implies that some coordinates of z must have a higher variance and lower reconstruction stability. Hence,

$$\mathbb{E}_z[\Delta E_{base}] \geq \beta D$$

for some $\beta > 0$ where $\beta > 0 \in \mathbb{R}$ and $\alpha \geq \beta$.

This holds because higher diversity requires the VAE to encode more varied information, necessarily increasing posterior uncertainty.

For a sampled latent vector $\mathbf{z} = \boldsymbol{\mu} + \boldsymbol{\sigma} \odot \boldsymbol{\epsilon}$ where $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, consider the telescoping sum:

$$\Delta\mathbb{E}_{\text{base}} = \mathbb{E}_{\text{sample}} - \mathbb{E}_{\text{mean}} \tag{2.10}$$

$$\leq \sum_{i=1}^{d}(\mathbb{E}_{\text{sample}} - \mathbb{E}_{\text{int},i}) \tag{2.11}$$

$$= \sum_{i=1}^{d} \Delta\mathbb{E}_i \tag{2.12}$$

The inequality becomes approximate equality when the decoder exhibits near-additive behavior with respect to latent perturbations while also maintaining the standard regularity conditions of continuity and smoothness. Under these conditions on the decoder, we have:

$$\sum_{i=1}^{d} R_i = \sum_{i=1}^{d} \frac{\Delta\mathbb{E}_i}{\Delta\mathbb{E}_{\text{base}}} \geq \alpha \cdot \gamma \tag{2.13}$$

where $\alpha \cdot \gamma \geq \beta \cdot \gamma$.

Suppose, for contradiction, that $R_i \leq \tau$ for **all** dimensions $i \in \{1, \ldots, d\}$.

Then:

$$\sum_{i=1}^{d} R_i \leq d \cdot \tau \tag{2.14}$$

If $\tau < \frac{\alpha\gamma}{d}$, then:

$$\sum_{i=1}^{d} R_i \leq d \cdot \tau < d \cdot \frac{\alpha\gamma}{d} = \alpha\gamma \tag{2.15}$$

This contradicts the inequality in Equation 10.

Therefore, there must exist at least one dimension $i^* \in \{1, \ldots, d\}$ such that $R_{i^*} > \tau$. □

3 Results

With the above methodology, the experiment was performed. The datasets were split into having 1 class, 5 classes and 10 classes, each having 1000 samples. For these datasets, the mean with standard error is given in tables 3.1 and 3.2.

Chunk	Diversity	Var_Div	Recon_Samp	Recon_Mean
0	0.662±0.039	0.022±0.002	0.029±0.003	0.027±0.003
1	0.781±0.007	0.020±0.001	0.017±0.0005	0.013±0.0004
2	0.796±0.002	0.016±0.0002	0.015±0.0002	0.011±0.0001
3	0.709±0.044	0.027±0.004	0.026±0.002	0.025±0.002
4	0.855±0.004	0.024±0.002	0.018±0.0005	0.015±0.0006
5	0.879±0.001	0.018±0.0003	0.017±0.0002	0.015±0.0002

Table 3.1: Summary of primary dataset and reconstruction metrics (Mean ± SE). Chunks 0–2 correspond to MNIST with 1, 5, and 10 classes, respectively. Chunks 3–5 correspond to FashionMNIST with 1, 5, and 10 classes.

Chunk	Recon_Diff	Ratio_Sum	Sig_Dims	Int_Var
0	0.002±0.0002	1.037±0.020	3.3±0.4	3.4e-09±2e-10
1	0.004±0.0001	0.974±0.005	0.1±0.1	2.0e-09±3e-10
2	0.004±0.0001	0.983±0.004	0.0±0.0	9.3e-10±4e-11
3	0.001±0.0004	0.391±0.846	7.6±2.1	7.2e-09±1e-09
4	0.003±0.0001	1.133±0.015	7.2±0.4	7.3e-09±5e-10
5	0.002±0.0001	1.019±0.006	6.0±0.6	5.9e-09±5e-10

Table 3.2: Summary of intervention-based uncertainty metrics (Mean ± SE). This table presents the key outcomes of our analysis. The FashionMNIST dataset can be found in (Chunks 3–5) and the MNIST in (Chunks 0–2).

Then, the Spearman Correlation Coefficients were calculated. Spearman was used over Pearson as it captures monotonic relationships and does not assume linear relationships with the variables. The spearman correlation was calculated between all the metrics and the significant dimensions. This was to see which metric was able to influence the significant dimensions of the dataset.

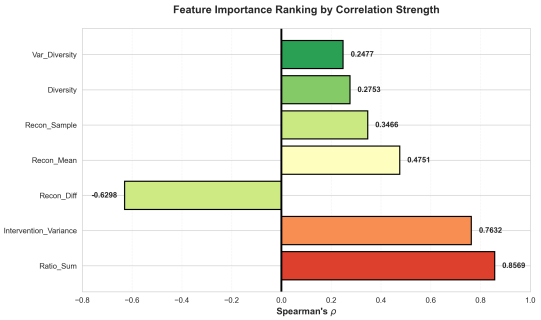

Figure 3.1: Visual ranking of feature importance based on Spearman's correlation strength with the number of significant dimensions.

Feature	Spearman's ρ	p-value
Ratio_Sum	0.8569	2.45×10^{-18}
Recon_Diff	−0.6298	7.02×10^{-8}
Recon_Mean	0.4751	1.25×10^{-4}
Recon_Sample	0.3466	6.67×10^{-3}
Diversity	0.2753	3.33×10^{-2}
Var_Diversity	0.2477	5.64×10^{-2}

Table 3.3: Spearman's rank correlation coefficients with the number of significant dimensions. The results show that `Ratio_Sum` has the strongest positive correlation ($\rho = 0.8569$, $p < 0.001$).

From the table 3.3, we can clearly see that the most significant factor in influencing the number of Significant dimensions is the Ratio_Sum with a $\rho = 0.8569$ and $p = 2.45 \times 10^{-18}$. With this, we get $\rho^2 = 0.7343$, explaining 73.43% of the variance in the number of Significant Dimensions. Figure 3.1 ranks the features according to their Spearman Coefficients.

Figure 3.2 shows the number of significant dimensions per dataset. Fashion MNIST has a greater number of mean significant dimensions per dataset, which corresponds also to the higher reconstruction sum tabulated above. It also has significantly higher standard error,

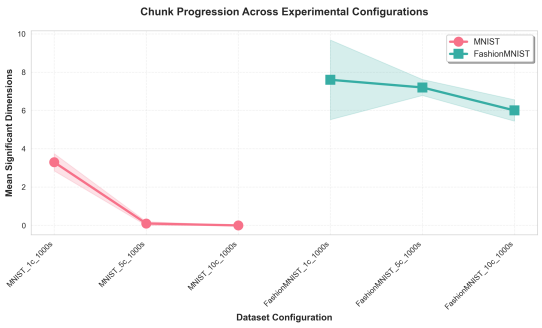

Figure 3.2: Mean number of significant latent dimensions across the experimental configurations. The shaded regions represent the standard error across the independent runs for each configuration.

showing greater variance in the data. Clearly, the mean number of significant dimensions per dataset in MNIST is much lower.

4 Discussion

Our investigation establishes a compelling, albeit complex, relationship between training data diversity and the structure of latent uncertainty in VAEs. The primary finding is that, generally, datasets with higher visual diversity tend to produce VAEs with a more structured latent space, where significant reconstruction uncertainty (set by an arbitrary ratio of 0.075 of the total ratio sum) can be attributed to a larger number of distinct dimensions. This observation offers us insights for developing more efficient, post-hoc methods for model refinement.

Our results partially support our central hypothesis, given the low Spearman's p. The core of our hypothesis (that higher diversity enables more effective interventions) is backed by the significant correlation found between our diversity metric and the number of significant dimensions (`Nsig`), as shown in Table 3.3, albeit with a low Spearman's p. The strong performance of the `Ratio_Sum` metric further suggests that with diverse data, the VAE is incentivized to learn a more disentangled representation, as the total uncertainty can be more effectively decomposed into the sum of its parts. Yet, the variability within our results is also telling. As noted in Section 4.1, some runs on the lower-diversity MNIST dataset yielded a surprisingly high count of significant dimensions, occasionally outperforming FashionMNIST runs. This suggests that while diversity is a powerful driving factor, the relationship is not perfectly deterministic. The stochasticity of both model training given different datasets and, critically, the random sampling of classes and images for each experimental run can introduce significant sample variance in the results. A low-diversity dataset by our average measure might, by chance, contain a few highly distinct samples that force the model to learn a more complex representation than is typical for that diversity level.

The underlying mechanism for our main finding can be traced back to the VAE's objective function. A VAE must balance the competing pressures of accurate reconstruction and regularization via the KL divergence term. When trained on a homogeneous dataset, the model can achieve low reconstruction error by mapping a narrow range of inputs to a small, tightly clustered region of the latent space. There is little incentive to utilize the full latent capacity. Conversely, a diverse dataset presents the model with a wider variety of features it must learn to reconstruct. To do so effectively, the encoder is forced to map different inputs to more distinct regions of the latent space, effectively "stretching out" the representation and encouraging the use of multiple dimensions to capture different factors of variation. This process naturally leads to a more structured and interpretable latent space where interventions are more meaningful.

4.1 Limitations

While our controlled diversity framework provides a reproducible way to study the role of dataset variety in epistemic uncertainty, several limitations remain.

First, the significance threshold ($\tau = 0.075$) used to identify active latent dimensions is arbitrary. Although it provides a simple heuristic for comparing across runs, its fixed nature may

over- or under-estimate the number of meaningful dimensions depending on the dataset or scale of the reconstruction errors.

Furthermore, it is crucial to critically evaluate our methodological choice for quantifying diversity. While using an aggregated SSIM score proved to be an effective proxy for diversity—as evidenced by the strong final correlations—it is not without significant practical drawbacks. The pairwise nature of SSIM results in a quadratic computational complexity ($O(n^2)$), making its application to the full training set computationally infeasible. Our decision to estimate diversity from a subset of 200 samples was a necessary compromise. This introduces a potential source of sampling error; our diversity score, D, is an estimate, not a ground truth. This limitation means we are correlating our intervention metrics with a potentially noisy measure of diversity, which may partly explain the observed variance in our results and temper the strength of our conclusions. A more scalable, globally-aware diversity metric could potentially reveal an even stronger, more stable relationship.

Third, the variance across random dataset samples can introduce inconsistencies in the observed metrics. Despite using identical architectures and fixed training seeds, some MNIST runs showed much higher counts of significant latent dimensions and larger ratio sums, even outperforming FashionMNIST. This variability suggests that the random selection of classes and samples can have a meaningful impact on the uncertainty measures, even within the same nominal configuration.

Furthermore, the reliance on the reconstruction error for identifying active latent dimensions could conflate the representation quality with the true dimensional usage. For example, a latent dimension with low variance in reconstruction error could either indicate actual inactivity or could just be an indicator of consistent encoding across the dataset. On the other hand, high variance could be due to the model's difficulty of reconstructing specific features rather than a meaningful usage of the dimension. This makes it difficult to distinguish between dimensions that capture important information and those that just reflect model instability or reconstruction noise.

Finally, the experiments were limited to grayscale datasets (MNIST and FashionMNIST) and a single, simple VAE architecture. These constraints restrict how far the conclusions can generalize to more complex data distributions.

4.2 Future Work

Future research could address these limitations and extend the current findings in several directions.

A natural next step would be to introduce adaptive or data-driven thresholds for identifying significant latent dimensions, perhaps using confidence intervals or bootstrapped variability measures. One could then potentially intervene on only the most significant dimensions, eliminating most of the reconstruction error while simultaneously keeping reconstructions varied.

Testing on larger or more structured datasets (e.g., CIFAR-10, CelebA, or medical imaging data) would also be valuable, as it could reveal whether the same relationships and trends hold for richer feature distributions.

In case these datasets give some more stable result, one could investigate how the latent space of less and more varied datasets evolve over time when training progresses. Specifically, tracking active latent dimensions over epochs could reveal whether high-diversity datasets activate more dimensions at an early stage and how this evolves over time.

5 Conclusions

The experiment reveals a complex relationship between training data diversity and VAE latent space structure. Our results weakly support our hypothesis: diversity correlates with significant dimensions, though the weak p-value indicates a non-deterministic relationship. `Ratio_Sum`'s strong predictive power suggests diverse data incentivizes disentangled representations. The mechanism arises from the VAE objective function: homogeneous datasets achieve low recon-

struction error with tightly clustered latent regions. Diverse datasets force the encoder to map inputs across more distinct latent regions, effectively stretching the representation across multiple dimensions to capture distinct factors of variation. This naturally produces structured, interpretable latent spaces amenable to meaningful interventions, though individual instances may deviate from aggregate patterns.

6 Reflections & Contributions

Our group adopted a hybrid of "Centralized and hierarchical system" together with the "Collegial system" throughout the project, mostly implicitly. We did not have internal conflicts, although the task division could have been done more clearly and fairly from the start. Each member's contribution was as follows:

- **Alexander**: Main conceptual development of research trajectory and methodology, did the code necessary for the experiment and wrote the methodology. Minor editing and language checks on introduction, results, and conclusion. Wrote the discussion, appendices on convolutional VAE and the old methodology.
- **Mika**: Wrote introduction. Minor editing and language checks on introduction, results, discussion, conclusion, and appendices. Wrote section B.2 in appendices.
- **Tim**: Wrote theoretical framework (both in main paper and appendix). Minor editing and language checks on introduction, results, discussion, conclusion, and appendices.
- **Aalam**: Wrote theoretical framework (both in main paper and appendix). Minor editing and language checks on introduction, results, discussion, conclusion, and appendices. Performed the statistical tests on the final results to obtain the Spearman's coefficients. Made figures and tables for results section. Wrote Results and Conclusion.
- **Adrian**: Minor editing and language checks on introduction, results, discussion, conclusion, and appendices. Wrote abstract, hyperparameters section in appendices, initial drafts of limitations and future work sections. Made heavier updates on methodology sections 2.1 and 2.2 to reflect the new code implementation with training subsets and SSIM diversity.

References

[1] Dr. Miguel Sanchez. Variational autoencoders - theory and applications: Exploring variational autoencoder models and their applications in generative modeling, representation learning, and beyond. *Advances in Deep Learning Techniques*, 4(1):18–32, Feb. 2024.

[2] BytePlus. Benefits of variational autoencoders (vae). https://www.byteplus.com/en/topic/400754?title=benefits-of-variational-autoencoders-vae, 2025. Accessed: 2025-10-07.

[3] Jianing Liu. Research on the application of variational autoencoder in image generation. *ITM Web of Conferences*, 70, 01 2025.

[4] Szilárd Molnár and Levente Tamás. Variational autoencoders for 3d data processing. *Artif. Intell. Rev.*, 57(2):42, February 2024.

[5] IBM. What is a variational autoencoder? https://www.ibm.com/think/topics/variational-autoencoder, 2025. Accessed: 2025-10-07.

[6] Talbot West. What are variational autoencoders and how are they useful? https://talbotwest.com/ai-insights/what-is-a-variational-auto-encoder-vae, 2025. Accessed: 2025-10-07.

[7] Towards Data Science. Uncovering anomalies with variational autoencoders (vae): A deep dive into the world of anomaly detection. https://towardsdatascience.com/uncovering-anomalies-with-variational-autoencoders-vae-a-deep-dive-into-the-world-of-1b2bce47e2e9/, 2025. Accessed: 2025-10-07.

[8] K. B. Anusha, Modalavalasa Divya, K. Madhuri Pravallikha Rani, B. Satvika, P. Tarun, G. Vaishnavi, and S. Linga Raju. Molecule generation of drugs using vae. In *Proceedings of the International Conference on Computational Innovations and Emerging Trends (ICCIET- 2024)*, pages 170–179. Atlantis Press, 2024.

[9] Aman Singh and Tokunbo Ogunfunmi. An overview of variational autoencoders for source separation, finance, and bio-signal applications. *Entropy*, 24(1):55, 2021.

[10] Jannes Klaas. *Machine Learning for Finance: Principles and Practice for Financial Insiders*. Packt Publishing, 2019.

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

[12] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. *arXiv preprint arXiv:1401.4082*, 2014.

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

[14] Carl Doersch. Tutorial on variational autoencoders. *arXiv preprint arXiv:1606.05908*, 2016.

[15] Gustav Bredell, Kyriakos Flouris, Krishna Chaitanya, Ertunc Erdil, and Ender Konukoglu. Explicitly minimizing the blur error of variational autoencoders, 2023.

[16] Yixin Wang, David M. Blei, and John P. Cunningham. Posterior collapse and latent variable non-identifiability, 2023.

[17] Diederik P. Kingma and Max Welling. An introduction to variational autoencoders. *Foundations and Trends in Machine Learning*, 12(4):307–392, 2019.

[18] Max. Why sampling in vae? https://medium.com/thedeephub/why-sampling-in-vae-2a3806de14ee, January 28 2024. Accessed: 2025-10-16.

[19] Variational autoencoders (vaes): A gentle guide to probabilistic generative models. https://ai.plainenglish.io/variational-autoencoders-vaes-a-gentle-guide-to-probabilistic-generative-models-8e53da5b71b7, 2025. Accessed: 2025-10-16.

[20] Dmitrii Matveichev. How to sample from latent space with variational autoencoder. https://hackernoon.com/how-to-sample-from-latent-space-with-variational-autoencoder, 2024. Accessed: 2025-10-16.

[21] DiShi Zhu. Generate images using variational autoencoder (vae). https://medium.com/@judyyes10/generate-images-using-variational-autoencoder-vae-4d429d9bdb5, April 19 2020. Accessed: 2025-10-16.

[22] Shivam Shinde. Image generation using vaes. https://www.kaggle.com/code/shivamshinde123/image-generation-using-vaes, 2025. Accessed: 2025-10-16.

[23] Infermatic.ai. What are the trade-offs between reconstruction loss and kl divergence in vaes? https://infermatic.ai/ask/?question=What+are+the+trade-offs+between+reconstruction+loss+and+KL+divergence+in+VAEs%3F, 2025. Accessed: 2025-10-16.

[24] Edureka. What is variational autoencoder architecture? a full guide. https://www.edureka.co/blog/variational-autoencoder-architecture/, 2025. Accessed: 2025-10-16.

[25] Andrea Asperti and Matteo Trentin. Balancing reconstruction error and kullback-leibler divergence in variational autoencoders. *IEEE Access*, 8:199440–199447, 2020. Open Access under CC BY 4.0 License.

[26] Vae loss function: Reconstruction & kl divergence. https://apxml.com/courses/applied-autoencoders-feature-extraction/chapter-6-variational-autoencoders-structured-latent-spaces/vae-loss-function-reconstruction-kl-divergence, 2025. Accessed: 2025-10-16.

[27] Tingsong Ou. Variational autoencoder, and a bit kl divergence, with pytorch. https://medium.com/@outerrencedl/variational-autoencoder-and-a-bit-kl-divergence-with-pytorch-ce04fd55d0d7, December 31 2022. Accessed: 2025-10-16.

[28] Rushikesh Shende. Autoencoders, variational autoencoders (vae) and β-vae. https://medium.com/@rushikesh.shende/autoencoders-variational-autoencoders-vae-and-%CE%B2-vae-ceba9998773d, April 19 2023. Accessed: 2025-10-16.

[29] Lucas Bechberger. What is a β variational autoencoder? https://lucas-bechberger.de/2018/12/07/what-is-a-%CE%B2-variational-autoencoder/, December 7 2018. Accessed: 2025-10-16.

[30] Arshid. Iris flower dataset. https://www.kaggle.com/datasets/arshid/iris-flower-dataset, 2018. Accessed: 2025-10-26; Originally from R. A. Fisher, 1936, "The use of multiple measurements in taxonomic problems", UCI Machine Learning Repository.

[31] Ron Edgar, Michael Domrachev, and Alex E. Lash. Gene expression omnibus: Ncbi gene expression and hybridization array data repository. *Nucleic Acids Research*, 30(1):207–210, 2002.

[32] Minjae Lee et al. Data-informed model complexity metric for optimizing model selection. *arXiv preprint arXiv:2501.17372*, 2025.

[33] Yael Mathov, Eden Levy, Ziv Katzir, Asaf Shabtai, and Yuval Elovici. Not all datasets are born equal: On heterogeneous tabular data and adversarial examples. *Information Sciences*, 588:192–208, 2022.

[34] Yaqing Shen and Yuxin Qin. Regularizing variational autoencoder with diversity and uncertainty awareness. *arXiv preprint arXiv:2303.11211*, 2024.

[35] Dora Zhao, Jerone T. A. Andrews, Orestis Papakyriakopoulos, and Alice Xiang. Position: Measure dataset diversity, don't just claim it. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024.

[36] Brando Miranda, Alycia Lee, Sudharsan Sundar, Allison Casasola, Rylan Schaeffer, Elyas Obbad, and Sanmi Koyejo. Beyond scale: the diversity coefficient as a data quality metric demonstrates llms are pre-trained on formally diverse data. In *Data-centric Machine Learning Research Workshop, International Conference on Learning Representations (ICLR)*, 2023.

[37] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *International conference on learning representations*, 2017.

[38] Alexey Dosovitskiy and Thomas Brox. Generating images with perceptual similarity metrics based on deep networks. *Advances in neural information processing systems*, 29, 2016.

[39] Emma Strubell, Ananya Ganesh, and Andrew McCallum. Energy and policy considerations for deep learning in nlp. *arXiv preprint arXiv:1906.02243*, 2019.

[40] David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. Network dissection: Quantifying interpretability of deep visual representations. In *Pro-*

ceedings of the IEEE conference on computer vision and pattern recognition, pages 6541–6549, 2017.

[41] 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.

[42] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. In *arXiv preprint arXiv:1708.07747*, 2017.

[43] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

[44] Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[45] Tarin Clanuwat, Marcel Bober-Irizar, Asanobu Kitamoto, Alex Lamb, Kazuaki Yamamoto, and David Ha. Deep learning for classical japanese literature. In *arXiv preprint arXiv:1812.01718*, 2018.

[46] Gregory Cohen, Saeed Afshar, Jonathan Tapson, and Andre Van Schaik. Emnist: Extending mnist to handwritten letters. *2017 International Joint Conference on Neural Networks (IJCNN)*, pages 2921–2926, 2017.

[47] Abhinav Yadav, Rahul Kuvre, and Li Deng. The qmnist dataset. *arXiv preprint arXiv:1912.12942*, 2019.

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

[49] Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2nd edition, 2009.

A Hyperparameters

A.1 Model Architecture & Training

Table A.1: Model and training hyperparameters.

Parameter	Value
Model type	VAE
Latent dimension	20
Encoder layers	$784 \to 400 \to 20$
Decoder layers	$20 \to 400 \to 784$
Activation function	ReLU
Output activation	Sigmoid
Optimizer	Adam
Learning rate	1×10^{-3}
Batch size	256
Epochs	40
Loss function	BCE + KL divergence ($\beta = 1$)
Learning rate scheduler	Disabled
Training seed	2810 (fixed)

A.2 Experimental Configuration

Table A.2: Experimental setup hyperparameters.

Parameter	Value
Base datasets	MNIST, FashionMNIST
Input transform	ToTensor()
Subset configurations	1, 5, and 10 classes × 1000 samples/class
Runs per configuration	10 (independent seeds)
Dataset seeds	2810 – 2819 (varied)
Training seed	2810 (fixed)
Diversity samples	200
Intervention samples	500
Latent significance threshold	0.075

B Previous Methodology Approach

B.1 Experimental Design Overview

To systematically investigate the relationship between dataset characteristics and the manifestation of uncertainty in VAE reconstructions, we designed a multi-dataset experimental framework. Our approach consists of three key components:

1. Quantitative measurement of dataset diversity

2. Training of VAE models across multiple datasets

3. Intervention-based analysis of latent uncertainty

By examining nine distinct grayscale image datasets, we aim to understand whether and how dataset properties influence the sources of reconstruction blur in VAEs.

B.2 Dataset Selection and Diversity Measurement

We selected nine grayscale image datasets to ensure sufficient variability in visual complexity and content diversity:

- **MNIST** [41]: A cornerstone of machine learning, the MNIST (Modified National Institute of Standards and Technology) dataset is a collection of 70,000 handwritten digits, from 0 to 9. It is divided into 60,000 images for training and 10,000 for testing. Each image is a 28x28 pixel grayscale representation. This dataset was created by "re-mixing" samples from NIST's original datasets to better suit machine learning experiments. It has become a fundamental benchmark for image processing and classification algorithms.

- **FashionMNIST** [42]: Created as a more challenging drop-in replacement for the original MNIST dataset, FashionMNIST features 70,000 grayscale images of fashion products from 10 different categories. The dataset, provided by Zalando, is also split into 60,000 training and 10,000 testing images, each at a 28x28 pixel resolution. It serves as a benchmark for modern machine learning algorithms, offering a greater challenge than the classic handwritten digits.

- **KMNIST** [45]: This dataset features handwritten Japanese cursive characters known as Kuzushiji. It contains a total of 70,000 grayscale images, with 60,000 allocated for training and 10,000 for testing. Each image is 28x28 pixels and belongs to one of 10 classes.

- **EMNIST-Balanced** [46]: An extension of the original MNIST dataset, EMNIST-Balanced provides a more challenging and balanced set of handwritten characters. It includes 131,600 grayscale images across 47 balanced classes, with 112,800 for training and 18,800 for testing. The images are 28x28 pixels.

- **EMNIST-Letters** [46]: This dataset is a subset of the broader EMNIST collection, focusing specifically on handwritten English letters. It consists of 145,600 grayscale images distributed across 26 balanced classes, with 88,800 in the training set and a test set that varies in reported numbers but is around 14,800. The image resolution is 28x28 pixels.

- **EMNIST-Digits** [46]: As a digit-centric part of the EMNIST datasets, this collection contains 280,000 grayscale handwritten digits. The dataset is balanced across 10 classes, with 240,000 images for training and 40,000 for testing, each being 28x28 pixels.

- **QMNIST** [47]: This dataset is an expanded version of the classic MNIST. It includes the original 60,000 training images and restores the full 60,000 test images, a portion of which had been previously unreleased. All images are 28x28 pixel grayscale representations of handwritten digits.

- **Omniglot-Background** [48]: The Omniglot dataset is designed for developing more human-like learning algorithms and is split into background and evaluation sets. The background set contains 964 different handwritten characters from 30 distinct alphabets. Each character has 20 examples, all presented as 105x105 pixel black and white images.

- **Omniglot-Evaluation**[48]: Serving as the evaluation counterpart to the background set, this dataset includes 659 handwritten characters from 20 different alphabets, intended for testing the one-shot learning capabilities of models. Similar to the background set, there are 20 images for each character, each with a resolution of 105x105 pixels.

To quantify the intrinsic diversity of each dataset, we employ a feature-based diversity metric using pre-trained deep representations. Specifically, we extract embeddings from a ResNet-18 model pre-trained on ImageNet, removing the final classification layer to obtain 512-dimensional feature vectors. For each dataset, we:

1. Sample $N = 1000$ images uniformly at random from the training set

2. Convert grayscale images to RGB by replicating across three channels

3. Resize images to 224×224 pixels and apply ImageNet normalization:

$$\tilde{x} = \frac{x - \mu_{\text{ImageNet}}}{\sigma_{\text{ImageNet}}}$$

   where $\mu_{\text{ImageNet}} = [0.485, 0.456, 0.406]$ and $\sigma_{\text{ImageNet}} = [0.229, 0.224, 0.225]$

4. Extract the penultimate layer activations as embedding vectors $\mathbf{e}_i \in \mathbb{R}^{512}$

5. Compute pairwise cosine distances between all embedding pairs:

$$d(\mathbf{e}_i, \mathbf{e}_j) = 1 - \frac{\mathbf{e}_i \cdot \mathbf{e}_j}{\|\mathbf{e}_i\| \|\mathbf{e}_j\|}$$

6. Calculate the dataset diversity score as the mean pairwise distance:

$$\text{Diversity} = \frac{2}{N(N-1)} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} d(\mathbf{e}_i, \mathbf{e}_j)$$

This metric provides a continuous measure of visual heterogeneity: higher values indicate greater diversity in the semantic content and visual features across samples, while lower values suggest more homogeneous datasets with limited variation.

C Convolutional Variational Autoencoder

To investigate whether the observed patterns found with the VAE described in Section 2.3 generalize beyond fully-connected architectures, we additionally train a convolutional variational autoencoder (ConvVAE) on the same datasets. Convolutional networks naturally exploit the spatial structure of images and have demonstrated superior performance on vision tasks. The ConvVAE model consists of an encoder with convolutional layers, reparameterization, and a decoder with transpose convolutional layers. The encoder processes the input image $\mathbf{x} \in [0,1]^{28\times28}$ through three convolutional blocks:

1. **Block 1**: Convolution with 32 filters of size 4×4, stride 2, and padding 1, followed by batch normalization and ReLU activation, producing feature maps of shape $(32, 14, 14)$.

2. **Block 2**: Convolution with 64 filters of size 4×4, stride 2, and padding 1, followed by batch normalization and ReLU activation, producing feature maps of shape $(64, 7, 7)$.

3. **Block 3**: Convolution with 128 filters of size 3×3, stride 1, and padding 1, followed by batch normalization and ReLU activation, producing feature maps of shape $(128, 7, 7)$.

The resulting feature maps are flattened to a 6272-dimensional vector ($128 \times 7 \times 7 = 6272$), which is then passed through two fully-connected layers to produce the approximate posterior parameters: mean $\boldsymbol{\mu} \in \mathbb{R}^{20}$ and log-variance $\log \boldsymbol{\sigma}^2 \in \mathbb{R}^{20}$. Latent samples are drawn using the reparameterization trick [11], identical to the fully-connected model.

The decoder begins by mapping the 20-dimensional latent code through a fully-connected layer to a 6272-dimensional vector, which is reshaped to $(128, 7, 7)$. This representation is then passed through three transpose convolutional blocks:

1. **Block 1**: Transpose convolution with 64 filters of size 4×4, stride 2, and padding 1, followed by batch normalization and ReLU activation, producing feature maps of shape $(64, 14, 14)$.

2. **Block 2**: Transpose convolution with 32 filters of size 4×4, stride 2, and padding 1, followed by batch normalization and ReLU activation, producing feature maps of shape $(32, 28, 28)$.

3. **Block 3**: Convolution with 1 filter of size 3×3 and padding 1, followed by sigmoid activation to produce the final reconstruction $\hat{\mathbf{x}} \in [0,1]^{28\times28}$.

The ConvVAE is trained using the same objective as the fully-connected model, maximizing the evidence lower bound (ELBO): with binary cross-entropy reconstruction loss and the same KL divergence regularization. We employ the Adam optimizer [44] with learning rate 10^{-3}, batch size 256, and train for 10 epochs. Batch normalization layers are used throughout the encoder and decoder to stabilize training and improve convergence. All other training procedures, including the intervention analysis methodology (Section 2.4), remain identical to enable direct comparison between architectures. Unfortunately, as mentioned in the main

paper, due to complexity and time constraints, we were not able to properly run the experiment with the ConvVAE and present them in this report.

D Different Proof Approach

As mentioned in the paper, there have also been various attempts at building a proof that shows the existence of active latent dimension for a high-diversity data set. One of them will be shown in the following, which is partially based on the book by Pearl[49]:

Lemma 5 (Additive Decomposition under Independence)**.** *If the latent dimensions are approximately independent (which is an assumption that goes against what is happening in the actual VAE) then:*

$$\Delta E_{base} \approx \sum_{i=1}^{d} \Delta E_i \tag{D.1}$$

Proof. By the chain rule for interventions in Chapter 3.1 of Pearl (rewritten):

$$E_{\text{mean}} = E_{\text{sample}} - \sum_{i=1}^{d} \left[E_{\text{int},\{1,\dots,i-1\}}(\cdot) - E_{\text{int},\{1,\dots,i\}}(\cdot) \right] \tag{D.2}$$

where $E_{\text{int},S}(\cdot)$ denotes fixing all dimensions in S to their means.

Furthermore:

$$E_{\text{int},\{1,\dots,i\}}(\cdot) - E_{\text{int},\{1,\dots,i-1\}}(\cdot) \approx E_{\text{sample}} - E_{\text{int},i}, \tag{D.3}$$

and

$$\Delta E_{\text{base}} \approx \sum_{i=1}^{d} \Delta E_i. \tag{D.4}$$

Which also results in

$$\sum_{i=1}^{d} R_i \approx 1. \tag{D.5}$$

□

Furthermore, we construct a different Lemma:

Lemma 6 (Diversity Lower Bound)**.** *For a dataset with diversity D and variance $Var(D)$, there exists a constant $\alpha > 0$ (depending on Lipschitz constant L) such that:*

$$\Delta E_{base} \geq \alpha \cdot \left[D^2 + Var(D) \right] \tag{D.6}$$

Proof. By Taylor expansion of g around μ and the Lipschitz property ($\|g(z) - g(z')\| \leq L\|z - z'\|$):

$$\begin{aligned} \Delta E_{\text{base}} &= E_{\text{sample}} - E_{\text{mean}} && \text{(D.7)} \\ &= \mathbb{E}\left[\|g(\mu + \sigma \odot \varepsilon) - g(\mu)\|^2 \right] && \text{(D.8)} \\ &\geq \frac{L^2}{2} \cdot \mathbb{E}\left[\|\sigma\|^2 \right] && \text{(D.9)} \end{aligned}$$

From the definition of the diversity in 2.2

$$\mathbb{E}[\|\sigma\|^2] \geq \beta \cdot \left[D^2 + \text{Var}(D) \right] \tag{D.10}$$

for some constant $\beta > 0$.

Setting $\alpha = \beta L^2/2$ gives the result. □

We will now proceed with the main theorem proof:

Theorem 7 (Dataset Diversity Implies Significant Dimensions)**.** *Let d be the number of latent dimensions and $\tau = 1.5/d$ be the significance threshold. If*

$$D > \gamma \coloneqq \sqrt{\frac{\tau \cdot d}{\alpha}} \tag{D.11}$$

then there exists at least one dimension $i \in \{1, \dots, d\}$ such that $R_i > \tau$.

Proof. We proceed by contradiction. Assume for all $i \in \{1, \dots, d\}$:

$$R_i \leq \tau \tag{D.12}$$

By Lemma 5:

$$\sum_{i=1}^{d} R_i \approx 1 \tag{D.13}$$

Since $R_i \leq \tau$ for all i:

$$1 = \sum_{i=1}^{d} R_i \leq \sum_{i=1}^{d} \tau = d \cdot \tau \tag{D.14}$$

Also from $R_i = \Delta E_i / \Delta E_{\text{base}}$:

$$\Delta E_i = R_i \cdot \Delta E_{\text{base}} \leq \tau \cdot \Delta E_{\text{base}} \tag{D.15}$$

By Lemma 6, if $D > \gamma$:

$$\Delta E_{\text{base}} \geq \alpha \cdot D^2 \tag{D.16}$$
$$\geq \alpha \cdot \gamma^2 \tag{D.17}$$
$$= \alpha \cdot \frac{\tau \cdot d}{\alpha} \tag{D.18}$$
$$= \tau \cdot d \tag{D.19}$$

If all $R_i \leq \tau$ and $\sum_{i=1}^{d} R_i = 1$, then the contribution ratios are approximately uniform: $R_i \approx 1/d < \tau$ for all i.

For $\Delta E_{\text{base}} = \tau \cdot d$, if contributions were truly homogeneous:

$$\Delta E_i \approx \frac{\tau \cdot d}{d} = \tau \tag{D.20}$$

If $D > \gamma = \sqrt{\tau d/\alpha}$, then:

$$\Delta E_{\text{base}} \geq \tau \cdot d \tag{D.21}$$

For the sum $\sum_{i=1}^{d} \Delta E_i = \Delta E_{\text{base}} \geq \tau \cdot d$ to hold with all $\Delta E_i \leq \tau \cdot \Delta E_{\text{base}} = \tau^2 \cdot d$, we would need:

$$\tau \cdot d = \sum_{i=1}^{d} \Delta E_i \leq d \cdot (\tau \cdot \Delta E_{\text{base}}) = d \cdot \tau^2 \cdot d = \tau^2 d^2 \tag{D.22}$$

If $\text{Var}(R_i) > 0$, then not all R_i can equal $1/d$. Since $\sum R_i = 1$ and the mean is $1/d$, at least one R_i must exceed the mean.

For $\tau = 1.5 \times (1/d)$, at least one $R_i > 1/d$ implies at least one R_i can exceed τ when the variance is sufficiently large, which occurs when $D > \gamma$.

Therefore, our assumption that all $R_i \leq \tau$ must be false, and there exists at least one dimension i such that $R_i > \tau$. □