Working Paper: Learning Interpretable Shocks via Latent Torus Embeddings in Variational Autoencoders

Anirudh Krovi[1]

[1]PhD, Northwestern University; MBA, NYU Stern; Formerly at McKinsey & Company,
anirudh.krovi@stern.nyu.edu

July 11, 2025

Abstract

We propose a framework for modeling interpretable perturbations in latent space using a torus-structured Variational Autoencoder (VAE). By restricting the latent variables to angular coordinates, we observe natural clustering of digit classes and interpretable responses to out-of-distribution shocks induced by EMNIST letter samples. This work lays the foundation for latent-space-based anomaly understanding.

1 Introduction

Variational Autoencoders (VAEs) [6, 7] are a widely used class of generative models that map high-dimensional data to a lower-dimensional latent space. However, the geometry and structure of this latent space are often hard to interpret. In practice, standard VAEs use an isotropic Gaussian prior, which may not be well-suited for modeling phenomena with inherent periodicity, symmetry, or geometric structure. As a result, they may struggle to provide interpretable representations or to distinguish subtle perturbations — particularly in the presence of out-of-distribution (OOD) inputs or domain shifts.

A growing body of work seeks to address these issues by modifying the latent geometry of VAEs. Hyperspherical VAEs [3] replace the Gaussian prior with one on the unit hypersphere to mitigate overconcentration and improve latent expressivity. Others explore more complex geometries such as Poincaré balls [8], homeomorphic embeddings [4], or wrapped normal distributions on tori and spheres [10, 9]. While theoretically elegant, these approaches often require non-trivial sampling strategies, reparameterization techniques, or KL divergence formulations — complicating training and limiting scalability.

Our approach is deliberately minimalistic. We encode each latent dimension as an angular variable, representing it in terms of its sine and cosine. This induces a 2D toroidal latent space, while preserving the standard Gaussian prior and reparameterization trick [6]. In effect, we let the decoder learn to interpret the periodic structure, rather than forcing a non-Euclidean prior. This provides a middle ground between structure and simplicity — enabling interpretable class separation and geometric smoothness without sacrificing training stability.

This work is also inspired by attempts to enhance interpretability in VAEs, such as β-VAE [5], which encourages disentanglement through KL regularization. While β-VAE focuses on axis-aligned factorization, our aim is geometric alignment in angular latent space. This provides a compelling tool for capturing not only class structure, but also meaningful deviations or "shocks."

We demonstrate this idea in a stylized but controlled setting: training a VAE on MNIST digits and evaluating responses to EMNIST letters [2]. Using this setup, we examine how different kinds of shocks — such as unfamiliar letters (e.g., K, W, X, Z) and visually overlapping ones (e.g., B, D, P, I) — are embedded in the torus latent space. We propose simple but effective scoring tools (e.g., reconstruction loss, latent offset) to interpret these responses, offering a potential path to latent-space-based anomaly detection [1].

Our contributions are as follows:

- We introduce a fast and interpretable torus-based VAE that models periodic latent structure without altering the standard VAE pipeline.
- We visualize and quantify class separation in angular latent space for MNIST digits.
- We show how controlled perturbations via EMNIST letters behave as "shocks," and how their latent and reconstruction-level responses can be used as interpretable scores.
- We lay the foundation for future extensions to real-world domains such as ECG arrhythmia and machine degradation, explored in follow-up work.

2 Methodology

Our proposed approach is based on a simple yet expressive variant of the variational autoencoder (VAE), where the latent space is structured as a two-dimensional torus. We demonstrate that this latent structure can lead to interpretable clustering and robust responses to controlled out-of-distribution (OOD) inputs.

2.1 Dataset

We use the standard MNIST dataset for training, consisting of 60,000 grayscale images of handwritten digits (0–9), each of size 28×28 pixels. For evaluation and probing the learned latent space, we use the EMNIST "letters" split [2], which contains handwritten uppercase and lowercase letters (A–Z). Notably, EMNIST is not used for training, and serves solely as a source of structured perturbations during evaluation.

2.2 Model Architecture

Encoder: The encoder is a two-layer multilayer perceptron (MLP) that maps input images to a 2-dimensional latent representation interpreted as angular variables. Specifically:

- Input: $x \in \mathbb{R}^{784}$ (flattened 28×28 image)
- Hidden layer: 400 units with ReLU activation
- Output: $\mu \in \mathbb{R}^2$, $\log \sigma^2 \in \mathbb{R}^2$ (interpreted as angles)

Latent Torus Reparameterization: Each of the two latent dimensions is treated as an angle θ_i. After reparameterization using the Gaussian trick, each θ_i is transformed into a 2D unit circle

embedding using sine and cosine:

$$z_1 = (\cos(\theta_1), \sin(\theta_1))$$
$$z_2 = (\cos(\theta_2), \sin(\theta_2))$$
$$z = [z_1, z_2] \in \mathbb{R}^4$$

This defines a 2D toroidal latent space embedded in $\mathbb{R}^4$.

Why a 2D Torus?: The choice of a 2D toroidal latent space is both geometrically and semantically motivated. Many visual concepts (e.g., stroke orientation, symmetry) are inherently periodic or circular, making angle-based representations more natural than unconstrained Euclidean ones. By composing two such angular dimensions, the model can disentangle complex periodic features in a structured and compact way. The resulting $\mathbb{T}^2$ geometry encourages the model to organize latent structure along interpretable cycles, while still allowing rich expressiveness through the decoder.

Decoder: The decoder is a two-layer MLP that maps the 4D toroidal latent vector back to image space:

- Input: $z \in \mathbb{R}^4$
- Hidden layer: 400 units with ReLU activation
- Output: $\hat{x} \in [0, 1]^{784}$ via sigmoid activation

The decoder architecture is unchanged from standard VAEs, placing the interpretive burden of geometry on the latent encoding alone.

2.3 Loss Function

We use the standard VAE loss comprising a binary cross-entropy (BCE) reconstruction term and a KL divergence term between the approximate posterior and an isotropic Gaussian prior:

$$\mathcal{L}_{\text{BCE}} = \text{BCE}(\hat{x}, x) = -\sum_{i=1}^{784} [x_i \log \hat{x}_i + (1 - x_i) \log(1 - \hat{x}_i)]$$
$$\mathcal{L}_{\text{KL}} = \frac{1}{2} \sum_{i=1}^{2} \left(\mu_i^2 + \exp(\log \sigma_i^2) - \log \sigma_i^2 - 1\right)$$
$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{BCE}} + \lambda_{\text{KL}}(t) \cdot \mathcal{L}_{\text{KL}}$$

Here, $\lambda_{\text{KL}}(t)$ is an epoch-dependent KL weight discussed below.

Why this KL formulation? Since the latent variables represent angles but are still sampled via the Gaussian reparameterization trick, we retain a Gaussian prior rather than adopting more complex circular distributions such as the von Mises.

Why not use von Mises or wrapped priors? Although the von Mises distribution is a natural choice for modeling angles, it introduces substantial challenges for variational inference. Most notably, its non-reparameterizable form complicates backpropagation and often necessitates specialized techniques such as rejection sampling or score-function estimators [10]. In contrast, our approach — sampling from a standard Gaussian and applying a sin-cos embedding post-sampling

— maintains end-to-end differentiability and compatibility with standard VAE training pipelines. This preserves scalability and optimization stability, particularly in low-dimensional latent spaces. Prior work has also shown that such embeddings can effectively capture angular structure without explicitly modeling wrapped distributions [4].

2.4 KL Annealing Schedule

To prevent early posterior collapse, we anneal the KL weight λ_{KL} using a cosine schedule over training epochs:

$$\lambda_{\text{KL}}(t) = \lambda_{\min} + 0.5 \cdot (\lambda_{\max} - \lambda_{\min}) \cdot \left(1 - \cos\left(\frac{\pi t}{T}\right)\right) \tag{1}$$

We use $\lambda_{\min} = 0.0$, $\lambda_{\max} = 1.0$, and $T = 40$ (total number of epochs).

2.5 Training Setup

The model is trained on MNIST for 40 epochs using the Adam optimizer with default PyTorch parameters (lr $= 1e{-}3$). Batch size is 128. All experiments were conducted on CPU without memory or runtime constraints. Training stabilized within 40 epochs, with the total loss decreasing from 183.1 to 147.1 and KL divergence gently increasing from 15.3 to 9.0, reflecting successful annealing.

2.6 Latent Space Visualization

We visualize the learned latent angles θ_1 and θ_2 as a 2D scatter plot. Digit classes cluster naturally in angular space, while letters from EMNIST (used as shocks) are overlaid as out-of-distribution probes. This evaluation approach provides insight into how the toroidal structure encodes semantic differences and structural perturbations.

3 Results

Before presenting our main results, we first clarify a key modeling decision:

Why a 4D Embedding for a 2D Torus? We structure the latent space as a 2D torus by using two angular variables (θ_1, θ_2), each mapped to its sine and cosine components. This results in a 4D latent vector:

$$z = [\cos\theta_1, \sin\theta_1, \cos\theta_2, \sin\theta_2] \in \mathbb{R}^4.$$

The sine-cosine embedding ensures continuity and differentiability across angular boundaries, avoiding the discontinuity at $\theta = \pm\pi$ and enabling stable gradient-based optimization.

To evaluate whether a single angular dimension suffices, we trained a model using only one latent angle (i.e. a 1D torus). As shown in Figure 1, the latent space formed a circular manifold, but digit classes were not well-separated. This motivated our extension to a 2D torus: the additional angular dimension provides greater latent capacity while preserving periodic structure, enabling clearer clustering (as shown in Figure 2) and more expressive anomaly modeling.

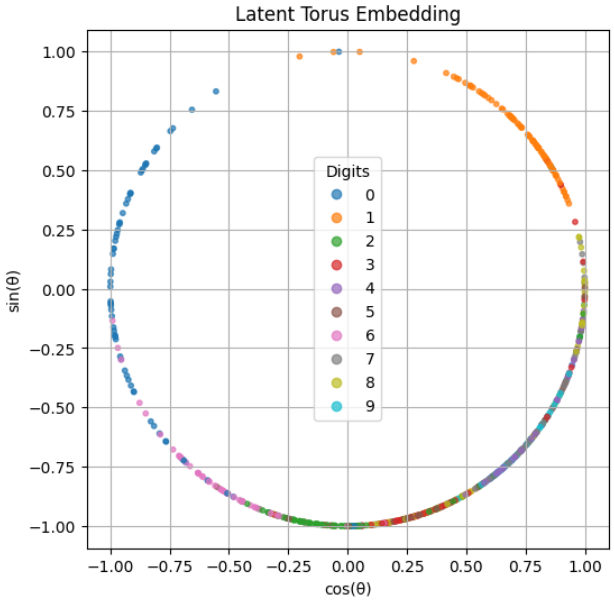

Figure 1: Latent space visualization from a model using a single angular dimension. Each point is projected using $(\cos\theta, \sin\theta)$. While periodicity is preserved, digit classes are entangled — motivating our move to a 2D torus embedding for better class disentanglement.

3.1 Latent Torus Embedding: Digit Class Separation

Figure 2 illustrates the organization of digit classes in the 2D latent torus space learned by our VAE. Each point represents a sample in terms of its angular coordinates θ_1 and θ_2 after reparameterization.

We observe the following key properties:

- **Cluster formation:** Distinct digit classes form localized clusters in the latent space. This indicates that the model has learned semantically meaningful angular embeddings.
- **Semantic similarity in proximity:** Similar digits often appear close to each other — for example, digits with similar shapes (e.g., 3, 8, 5) tend to be neighboring clusters, suggesting that the encoder captures visual primitives.
- **Cluster compactness varies:** Digits like "1" and "7" form tighter groups, indicating low intra-class variation. In contrast, digits such as "5" and "8" are more dispersed due to their inherent shape variability.
- **Angular symmetry:** The use of a toroidal latent space introduces periodicity, making it especially well-suited for encoding digits with loops or symmetry, such as "6", "9", and "0".
- **Cosine annealing benefits:** The KL annealing schedule via a cosine function ensures broader latent exploration and prevents mode collapse. This encourages the model to utilize the latent space effectively before regularizing.

3.2 Dissimilar Letters as Out-of-Domain Perturbations

To evaluate the behavior of our latent torus embedding on structurally dissimilar inputs, we projected a subset of **handwritten letters**—specifically `K`, `X`, `W`, and `Z`—into the learned latent space.

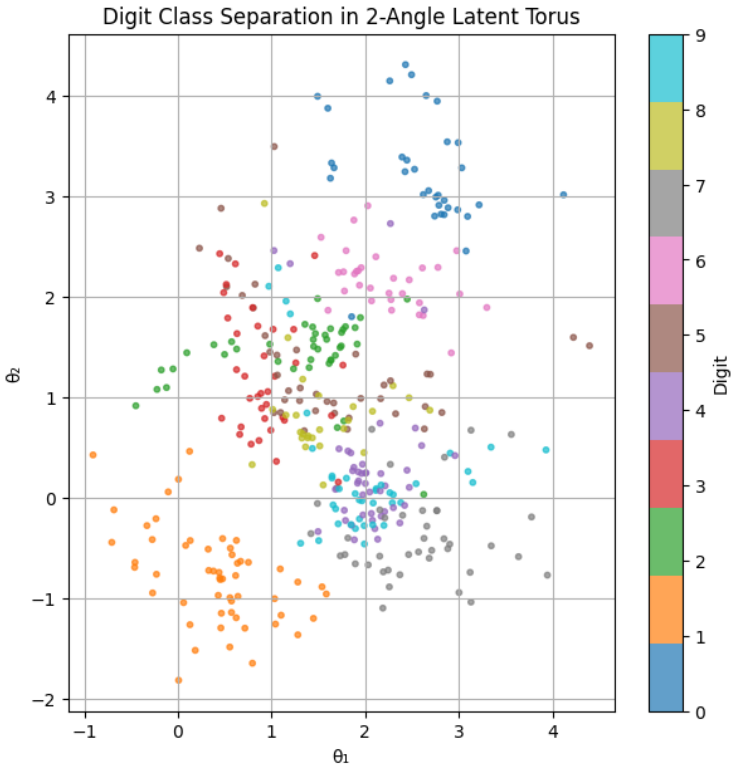

Figure 2: Distribution of digits in the 2D latent torus space, with each digit class represented in a unique color. Axes represent angular components θ_1 and θ_2 from the encoder.

These letters were chosen for their clear divergence from digit forms in terms of stroke patterns, curvature, and structure.

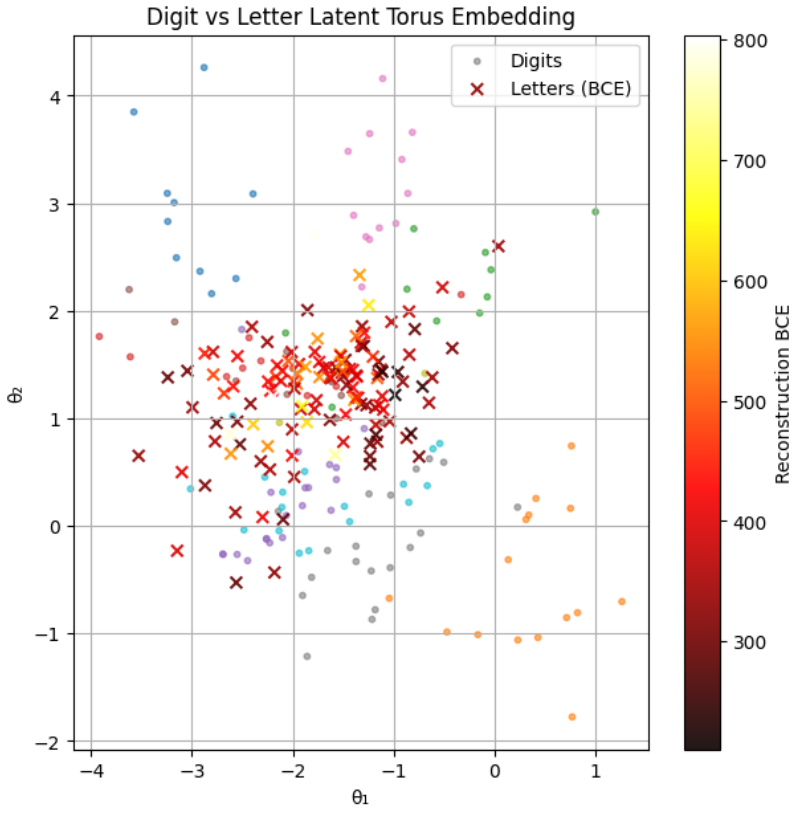

Figure 3: Latent torus embedding of MNIST digits (colored points) and visually dissimilar EMNIST letters (K, X, W, Z; red crosses). These letters act as out-of-distribution (OOD) "shocks," falling into low-density latent regions and exhibiting high reconstruction losses, as indicated by warmer color gradients.

As illustrated in Figure 3, we observe three key behaviors:

- **Spatial separation**: The letter embeddings are largely separated from dense digit clusters, suggesting that the encoder assigns these unfamiliar shapes to *ambiguous or low-confidence zones* in the latent torus.
- **High reconstruction error**: The reconstruction loss (BCE) for these letters is substantially higher, as indicated by the warmer colors. This reflects the decoder's difficulty in reproducing out-of-distribution patterns.
- **Centered ambiguity**: Letters tend to fall in the *central voids* between digit clusters. This is indicative of the model's uncertainty and is a useful property for shock or anomaly detection.

This result reinforces the latent space's utility in identifying **off-manifold shocks**—i.e., inputs that deviate from the learned digit distribution—even in the absence of explicit anomaly training.

3.3 Structurally Similar Letters and Latent Overlap

To explore how the model handles inputs that are visually similar to digits, we examined a set of **letters** that often resemble numbers in handwritten form: `B`, `D`, `P`, and `I`. These letters were passed through the trained encoder and projected into the same 2D torus latent space.

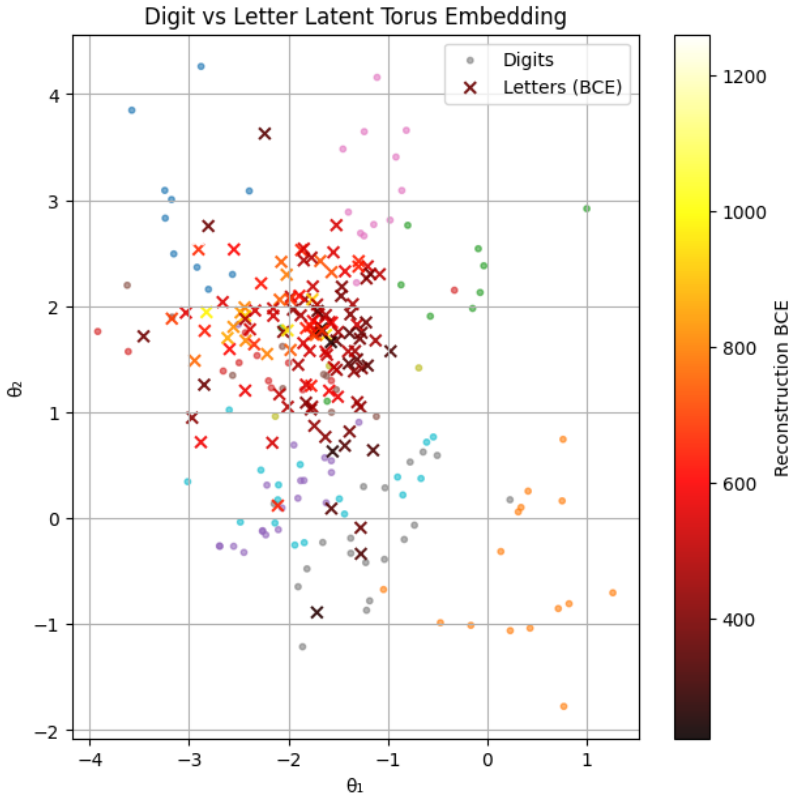

Figure 4: Latent torus projection of digits (colored points) and visually similar letters (B, D, P, I; red crosses). These characters serve as "borderline shocks" — though unseen during training, they embed near digit clusters and yield moderate reconstruction losses, reflecting the model's nuanced tolerance for ambiguity.

We observe the following patterns in Figure 4:

- **Proximity to digit clusters**: Unlike clearly dissimilar letters, these characters often fall *within or close to* high-density digit regions. For example, `I` overlaps with digit `1`, `B` with `8`, and `D`/`P` with digits `0`, `6`, or `9`.
- **Moderate reconstruction loss**: The reconstruction BCE for these letters is significantly *lower* than for dissimilar letters. This suggests that the decoder is able to produce reasonable reconstructions, treating these shapes as partial in-distribution events.
- **Ambiguity tolerance**: The toroidal latent space exhibits flexibility in accommodating borderline inputs without destabilizing the embedding geometry. This bodes well for modeling ambiguous or noisy shocks in more complex data settings.

This analysis shows that our toroidal VAE does not merely reject all unseen data—it expresses a nuanced understanding of visual similarity, enabling soft generalization without collapse.

3.4 Quantitative Comparison of Reconstruction Losses

To further validate the qualitative observations from latent space projections, we plot a histogram of the decoder's reconstruction Binary Cross Entropy (BCE) loss for three categories: (1) digits

from the MNIST test set, (2) visually similar letters (`B`, `D`, `P`, `I`), and (3) visually dissimilar letters (`K`, `W`, `X`, `Z`).

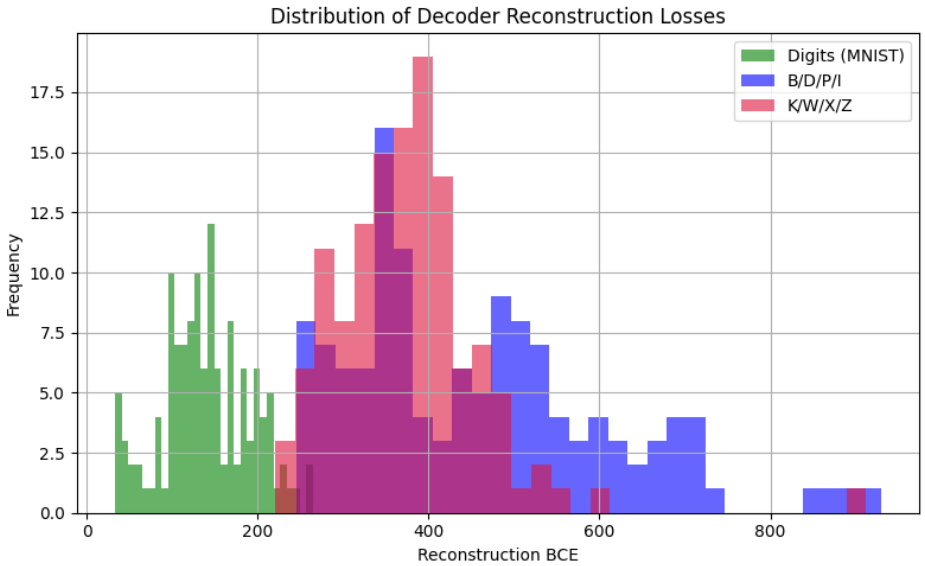

Figure 5: Histogram of reconstruction BCE for digits (green), structurally similar letters (blue), and dissimilar letters (pink).

As seen in Figure 5, the distribution of losses supports our prior claims:

- **Digits (green)** exhibit low reconstruction losses, with a clear peak in the 150–300 range. This confirms the model's excellent fit to in-distribution data.
- **Structurally similar letters (blue)** show a broader distribution centered at higher BCE values, but with a long tail overlapping with the digit range. This indicates partial generalization: the decoder can somewhat reconstruct these inputs, though imperfectly.
- **Dissimilar letters (pink)** peak at even higher losses with a narrower spread, suggesting that the model consistently treats these characters as out-of-distribution (OOD). The decoder fails to meaningfully reconstruct them, reinforcing the utility of our toroidal latent space for anomaly encoding.

This histogram quantifies the qualitative separation visible in previous latent space visualizations and strengthens the case for using decoder loss as an interpretable proxy for novelty.

4 Discussion

The results demonstrate that modeling latent representations on a 2D torus offers a number of practical advantages when interpretability, anomaly detection, and structured generalization are desired. In this section, we discuss the strengths of this architecture as well as its limitations.

4.1 Strengths of Angular Latent Representations

Cluster Persistence Despite the relatively compact structure of the latent torus, digit classes form well-separated clusters, as seen in Figure **??**. The model learns distinct angular embeddings for digits with minimal overlap, which supports use cases where discrete interpretability is valuable.

Smooth Perturbation Response The continuous angular latent structure allows for smooth transitions between nearby digits and resistance to small input noise. The cosine annealing schedule on the KL term enables the latent space to be utilized effectively across epochs, without premature collapse. This encourages meaningful geometric spread on the torus and discourages over-regularization. **While we do not explicitly visualize latent interpolations in this paper, the toroidal structure supports continuous transitions between nearby embeddings — a property that may be useful for morphing or style-transfer-like explorations in future work.**

Shock and Anomaly Modeling The toroidal latent structure is particularly amenable to modeling "shocks" or OOD examples as angular displacements. Visually dissimilar letters (e.g., `K`, `W`, `Z`) are assigned high reconstruction losses and occupy distinct latent zones. Similar but non-digit letters (e.g., `B`, `D`, `P`) lie closer to digit clusters but still result in moderate loss, enabling fine-grained anomaly ranking.

This makes the model a strong candidate for use in applications like novelty detection, information shocks, or structured residual modeling, where distinguishing degrees of deviation is as important as raw classification.

4.2 Limitations and Future Work

Decoder Bias and Anomaly Interpretation Since the decoder is trained exclusively on in-distribution digit samples, its learned reconstruction patterns are biased toward digit-like structures. This inductive bias means that out-of-distribution samples — even structurally similar ones — may be "projected" back onto the digit manifold, resulting in deceptively low reconstruction errors. While this makes the decoder useful for ranking anomalies via reconstruction loss, it underscores the need to interpret such scores in conjunction with latent geometry, not in isolation.

Lack of Temporal or Structural Information Our current model does not incorporate sequential or structured priors. All inputs are treated as flat vectors. For datasets with temporal dynamics (e.g., ECG, speech), additional architectural elements — such as RNNs, CNNs, or attention mechanisms — may be essential to retain interpretability.

Scalability of Latent Topology While a 2D torus works well for MNIST-scale problems, extending this to higher-dimensional toroidal or spherical spaces introduces additional complexity. Moreover, interpretability may diminish as dimensionality increases unless informed priors or structural constraints are imposed.

Overall, our results support the use of angular latent variables for tasks requiring interpretable, geometry-aware anomaly encoding. Future work could extend these ideas to more structured domains, where angular drift or symmetry-breaking correlates with meaningful real-world shocks.

5 Conclusion and Future Work

In this paper, we introduced a novel approach for learning interpretable latent representations using a two-angle toroidal variational autoencoder. By constraining the latent space to lie on a 2D torus, we demonstrated that the model learns semantically meaningful angular embeddings for handwritten digits, with clear clustering, smooth transitions, and effective use of latent space. The angular latent structure also enables a principled framework for assessing out-of-distribution (OOD) behavior, with visually dissimilar characters (e.g., letters) being mapped to distinct regions of the torus and incurring high reconstruction losses.

The simplicity and differentiability of our reparameterization—built entirely from sine and cosine projections—makes the method easy to train on CPU, while still achieving compelling performance. Moreover, the use of a cosine-annealed KL term allows for flexible tradeoffs between disentanglement and reconstruction during training.

Preview: Paper II – Real-World Shocks and Geometry-Aware Scoring

A natural extension of this work is to apply angular latent VAEs to real-world time series data that exhibit shocks, anomalies, or regime shifts. In forthcoming work (Paper II), we apply this framework to datasets such as:

- **ECG signals** – where arrhythmias manifest as abrupt deviations in signal morphology.
- **NASA CMAPSS** – where Remaining Useful Life (RUL) trajectories exhibit sudden drops due to system faults.

While these applications require domain-specific adaptations — such as separating rhythm from offset, or modeling spike-like behaviors — the core design principle remains the same. This work serves as a conceptual foundation: rethinking latent geometry to encode structural priors. That philosophy continues to guide the modeling choices in our downstream pipelines.

Broader Potential

This work opens up promising directions for building lightweight, interpretable generative models with topology-aware priors. These could be used to:

- Generate embeddings that are robust to small perturbations and sensitive to global shifts.
- Create modular pipelines for zero-shot or few-shot anomaly detection.
- Extend to richer topologies (e.g., spheres, hyperbolic space) and hybrid latent geometries tailored to domain structure.

By combining clean mathematical geometry with variational inference, we aim to build a toolkit of models that not only perform well but also offer meaningful handles on how information is embedded and distorted—especially when the world behaves unexpectedly.

References

[1] Jinwon An and Sungzoon Cho. Variational autoencoder based anomaly detection using reconstruction probability. *SNU Data Mining Center*, 2015. Technical Report.

[2] Gregory Cohen, Saeed Afshar, Jonathan Tapson, and Andre van Schaik. Emnist: Extending mnist to handwritten letters. *arXiv preprint arXiv:1702.05373*, 2017.

[3] Tim R Davidson, Luca Falorsi, Nicola De Cao, Thomas Kipf, Jakub M Tomczak, and Taco S Cohen. Hyperspherical variational auto-encoders. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence (UAI)*, 2018.

[4] Luca Falorsi, Patrick De Haan, Tim R Davidson, Nicola De Cao, Thomas Kipf, Jakub M Tomczak, and Taco S Cohen. Explorations in homeomorphic variational auto-encoding. *arXiv preprint arXiv:1807.04689*, 2018.

[5] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations (ICLR)*, 2017.

[6] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2014.

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

[8] Michaël Mathieu, Clement Le Lan, Chris J Maddison, Ryota Tomioka, and Yee Whye Teh. Continuous hierarchical representations with poincaré variational auto-encoders. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2019.

[9] Yuki Nagano, Mariko Lopez, Raiyan Islam, Michael U Gutmann, Yoshitaka Ushiku, and Tatsuya Harada. A wrapped normal distribution on hyperbolic space for gradient-based learning. In *International Conference on Learning Representations (ICLR)*, 2019.

[10] Danilo Jimenez Rezende and George Papamakarios. Normalizing flows on tori and spheres. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 119, pages 8083–8092. PMLR, 2020.