A Computation times for local descriptors

The computation times for the local scaling and local rank computation (since both require one randomized SVD computation for one latent vector) ends up being 3929s for 1000 samples. For local complexity we require 113s for 1000 samples. All the estimates are for a JAX implementation of Stable Diffusion on TPUv3.

Note that to train a reward model, we require the descriptors to be computed only once for each pre-trained model. If we compute the local scaling for 100k samples we require 173.1 TPU v3 hours which is equivalent to 54.58 V100 hours (according to Appendix A.3 Dhariwal & Nichol (2021)). Compared to 79,000 A100 hours required for Stable Diffusion training[2], 24000 hours with enterprise level optimization[3], the computation required for the descriptors and reward model training is significantly small. The computation time for the local descriptors can be further reduced by using a smaller k for our projection matrix W, or by using non-jacobian based methods, e.g., estimating the local scaling by measuring the change of volume for a unit norm ℓ_1-ball in the input space. We leave exploration of these directions for future work.

B Related Works

Local geometry pre-diffusion. Early applications of the local geometry of generative models involved improving the generation performance and/or utility of generative models via geometry inspired methods. For example, in Rifai et al. (2011) the authors proposed regularizing the contraction of the local geometry to learn better representations in autoencoders trained on MNIST and CIFAR10. The regularization penalty is employed via the norm of the input-output jacobian in Rifai et al. (2011), is an upper bound for local scaling presented in our paper. In Arvanitidis et al. (2017) the authors provided visualizations on the curvature of pre-trained VAE latent spaces and proposed using an auxiliary variance estimator neural network to regularize the latent space geometry during generation. In Kuhnel et al. (2018) the authors perform latent space statistical inference problems, e.g., maximum likelihood inference, by training a separate neural network to approximate the Riemannian metric. In Humayun et al. (2022a) the authors proposed a novel latent space sampling distribution based on the latent space geometry that allows uniformly sampling the learned data manifold of continuous-piecewise affine generators. The authors showed downstream benefits with fairness and diversity for such latent space samplers. While most of these methods discuss pre-diffusion architectures, their results are early demonstrations of how the local geometry can affect downstream generation. also employ auxiliary Neural Networks to model an intrinsic property of a pre-trained generator, similar to how we propose using a reward model for Stable Diffusion.

Local intrinsic dimensionality of diffusion models. The local geometry of diffusion models and possible applications have garnered significant interest in recent years. In **?** the authors propose a method to compute the intrinsic dimensionality of diffusion models using the assumption that the score field is perpendicular to the data manifold. For any vector x on the data manifold, the method requires computing the dimensionality of the score field around x and subtracting it from the ambient dimension. To do that, the authors perform one step of the forward diffusion process k times for x, denoise the k noisy samples using the diffusion model and compute the rank of the data matrix containing denoised samples to obtain the intrinsic dimensionality. Compared to this method, we compute the dimensionality directly via a random estimation of the input-output jacobian SVD. We do not require any assumption on the score function vector field being perpendicular to the data manifold, which may not hold for a diffusion model that is not optimally trained or highly complex training datasets like LAION.

In Kamkari et al. (2024a) the authors compute rank using the method proposed in **?** and show that local intrinsic dimensionality can be used for out-of-distribution (OOD) detection. This is analogous to our analysis in Sec 3 on the local geometry on or off the manifold. We can see that the intuition authors provided in Kamkari et al. (2024a) for diffusion models trained on smaller models and datasets e.g., FMNIST, MNIST, translate to larger scale models like Stable Diffusion trained on LAION as we have presented fig. 3, fig. 16 and Sec 4. Especially in section 4, we show that creating OOD samples with corruptions on Imagenet data (in-distribution), we can have an increase or decrease

[2]https://www.mosaicml.com/blog/training-stable-diffusion-from-scratch-costs-160k

[3]https://www.databricks.com/blog/stable-diffusion-2

in negative-log likelihood (estimated via local scaling), with decrease for blurring corruptions and increase in noising corruptions.

Concurrent work Kamkari et al. (2024b) has also shown the relationship between the intrinsic dimensionality (local rank) of Stable Diffusion scale models and the texture/visual complexity of generated images. We believe our analysis is much more holistic with three different geometric properties being measured compared to only local dimensinality. We i) show quantitatively how diversity measured via vendi score is higher for higher local scaling and rank values (section 4). We have explored how rank and scaling evolves continuously across the latent space in fig. 3. We have presented how the geometry distribution varies as we continually perturb images via noise or blurring operations section 4 And finally in Sec 5 we have presented a method to guide generation using the local geometry to obtain downstream generation benefits.

Misc. Apart from the aforementioned works, Kadkhodaie et al. (2023) show that the emergence of generalization in diffusion models – when two networks separately trained on the same data learn the same mapping – can be attributed to the eigenspectrum and eigenvectors of the input-output jacobian. While we do not study the training dynamics of the local geometric descriptors in our paper, Kadkhodaie et al. (2023) suggests that the local geometry can be an important indicator of memorization and generalization emergence in diffusion models. In Manor & Michaeli (2023) the authors use the posterior principal components of a denoiser for uncertainty quantification. This work suggests that components with larger eigenvalues result in larger uncertainty which is directly related to the local scaling descriptors as it measures the product of non-zero singular values. While in Manor & Michaeli (2023) the authors propose using it for only a single image denoiser, we show that it generalizes for any diffusion model including Stable Diffusion scale text-to-image diffusion models.

C Correlations between the three descriptors

Local scaling characterizes the change of volume by the affine slope $\mathbf{A}_\omega$ going from the latent space to the data manifold. *Local rank* characterizes the number of dimensions retained on the manifold after the network locally scales the latent space. Both local rank and scaling quantify first order properties of the CPWL operator. *Local complexity* approximates the 'number of unique affine maps' within a given neighborhood Humayun et al. (2024) by computing the number of CPWL knots intersecting an ℓ_1 ball in the input/latent space. Therefore local complexity is a measure of 'un-smoothness' and quantifies local second-order properties of a CPWL operator.

Correlations between local scaling ψ and local rank ν. By definition, local scaling and local rank are correlated, since both characterize the change of volume by the network input-output map at any input space linear region – also evident in eq. (2) and eq. (4). Local scaling is also upper bounded by local rank, $\psi_\omega \leq \sigma_0^{\nu_\omega}$ where σ_0 is the largest singular value of $\boldsymbol{A}_\omega$. The correlation is evident for our low dimensional DDPM setting presented in fig. 13, local rank and local scaling are highly correlated in their spatial distribution. There are indications suggesting that the correlations persists throughout training as can be seen in fig. 13 rightmost column top and bottom. However in fig. 3, we can see that in the high-dimensional Stable Diffusion latent space, local scaling and rank are correlated but local rank has sharper changes spatially compared to local scaling.

Correlations between local complexity δ and rank ν. There also exist correlations between local complexity and local rank due to the continuity of CPWL maps – between two neighboring linear regions ω_1 and ω_2, the corresponding slope matrices $\mathbf{A}_{\omega_1}$ and $\mathbf{A}_{\omega_2}$ differ by at most one row. Therefore between two neighboring regions ω_1 and ω_2, $|\nu_{\omega_1} - \nu_{\omega_2}| <= 1$. Informally, the local rank in a neighborhood V is lower bounded by the number of non-linearities in neighborhood V. This is evident in the empirical results presented in fig. 13 and fig. 3. In both figures, for input space neighborhoods with higher local complexity, we see a decrease in local rank. However, we do not observe sharp changes in local complexity as we observe in local rank in fig. 3. In fig. 13 we see that local rank is more discriminative of the data manifold compared to local complexity. Their training and denoising dynamics differ significantly as seen in fig. 13 rightmost column.

Qualitative and quantitative results on correlations. We train a beta-VAE unconditionally on MNIST and present in Fig. 9 samples from increasing local descriptor level sets from left to right along the columns. In Fig. 10, we present joint distributions of local scaling, complexity, rank and mean squared reconstruction error for training and test samples. We see that while local scaling, complexity and rank have some linear correlation, the classwise distribution in fig. 10 is very different between the three. We also present in fig. 11 the vendi score for increasing local scaling level sets and evidence that the population means for the descriptors don't follow the same pattern between sub-populations.

Figure 9: Joint distributions for local scaling and MSE, local rank and MSE, local complexity and MSE, local scaling and local complexity, and local scaling and local rank. We observe that local complexity is linearly correlated wth MSE, with higher complexity images incurring higher error. Local scaling, rank and complexity have correlations between them as well.

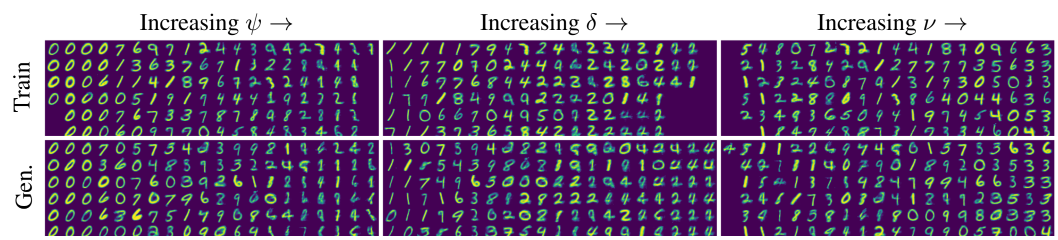

Figure 10: Level sets of data manifold descriptors for a Beta-VAE trained unconditionally on MNIST. From left to right, we present training samples (top row) and generated samples (bottom row) for linearly increasing level sets of local scaling (ψ) from $[-80, -42]$, local complexity (δ) from $[0, 120]$ and local rank (ν) from $[1.5, 5.5]$. Not all level sets had an equal number of samples from training/generated distributions. We see that for higher ψ, we have more outlier samples whereas for lower ψ we have modal samples. For increasing δ we see that the quality of generated samples decreases and the diversity of samples is reduced as well. For higher ν digits become more regularly shaped.

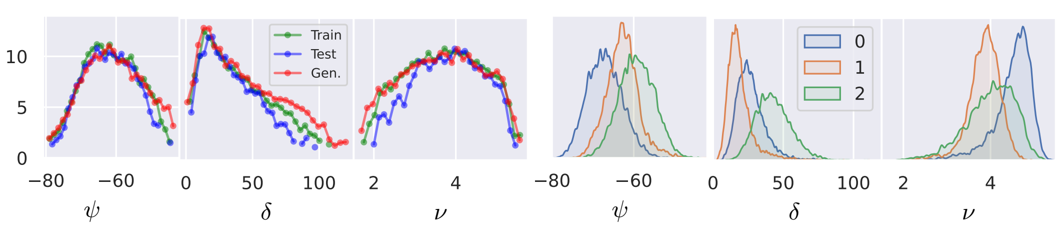

Figure 11: (Left panel) Vendi score (Friedman & Dieng, 2023) calculated for samples from different local descriptor level sets of a Beta-VAE. We take upto 150 samples from each level set and compute vendi score seperately for the MNIST train dataset, test dataset and generated samples. (Right panel) Sub-population differences of local descriptors in training data. We see that the order of sub-population means for the three classes, are not the same for all three descriptors.

D Additional Experiments

D.1 Local scaling for transformer based diffusion model

Since we are based on the CPWL formulation of NNs, our framework would generalize to models of any scale and any architecture with CPWL non-linearities. Empirically we have shown it to generalize for non-CPWL architectures like Stable Diffusion v1.4 and DDPM that employs non CPWL non-linearities such as attention, GeLU and much more. Following suggestions by the reviewer, we have performed additional experiments with a DiT-XL Peebles & Xie (2023) trained on Imagenet-256. For the DiT we compute the descriptors for the transformer network, conditioned on noise level $t = 0$, i.e., zero noise level. We generate 5120 images conditioned on Imagewoof Howard (2019) classes and present in fig. 12, increasing local scaling level sets from left to right. We see that similar to fig. 16 from the, DiT exhibits a qualitative correlation between visual complexity and local scaling. For additional analysis we repeat the Stable Diffusion experiments on the relation between diversity and local scaling for DiT. We see that similar to Stable Diffusion, for increasing local scaling level sets, the diversity of images increase and then drop for the highest local scaling level sets.

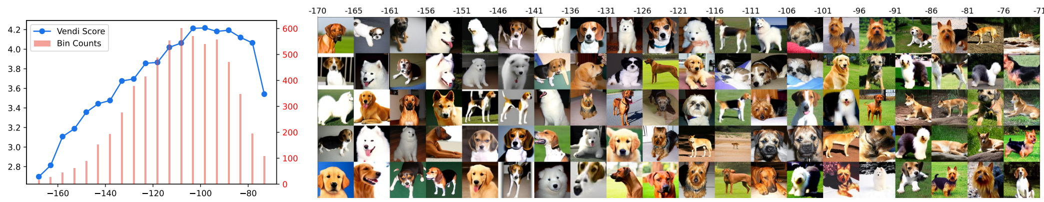

Figure 12: Left: Vendi score and membership counts for increasing local scaling level sets, computed for a DiT transformer. We see that similar to Stable Diffusion, local scaling increases from lower to higher local scaling level sets, then drops for very high local scaling level sets. Right: Generated samples from each level set in the left panel. Sample sets from higher local scaling level sets, tend to be more diverse.

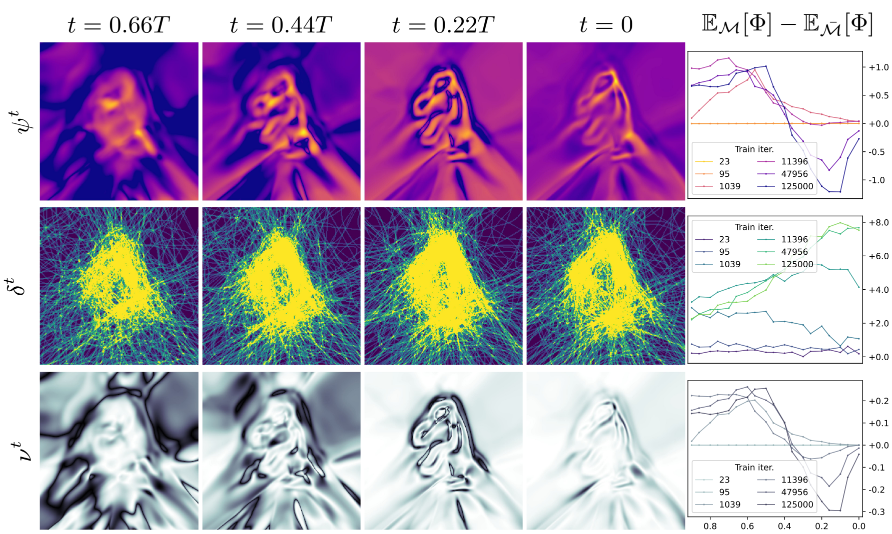

Figure 13: **Local geometric descriptors** computed over the input domain of a pre-trained toy diffusion model trained to produce samples from a dinosaur manifold $\mathcal{M} \in \mathbb{R}^2$. Descriptors are computed by conditioning the diffusion model on noise level t. We consider the set of input vectors within 0.05 units of the training data as on manifold $\mathcal{M}$ and rest as off the manifold $\bar{\mathcal{M}}$. We present the difference between the expected descriptor values on and off the manifold, $\mathbb{E}_{\mathcal{M}}[\Phi] - \mathbb{E}_{\bar{\mathcal{M}}}[\Phi], \forall \Phi \in \{\psi^t, \delta^t, \nu^t\}$ at different training iterations (right). We also present the descriptor computed over $[-6, 6]^2$ for different noise levels t after 125000 training iterations (rest). We observe that ψ^t is lower, δ^t is higher and ν^t is lower on the manifold than off the target manifold for lower noise levels, especially after the model is considerably trained. This indicates that for well trained diffusion model, i.e., learned manifold $\hat{\mathcal{M}} \approx \mathcal{M}$, local descriptors can distinguish between on and off manifold vectors in the input space.

D.2 VAE TRAINING DYNAMICS FOR MNIST

Setup. We train a Variational Auto Encoder (VAE) on the MNIST dataset with width 128 and depth 5 for both encoder and decoder. We add Gaussian noise with standard deviation $\{0, 0.0001, 0.001, 0.01, 0.1\}$ to the training data. Initialization was not kept fixed. In Fig. 14, we present plots showing the training dynamics of local complexity and scaling, averaged over all test dataset points from MNIST.

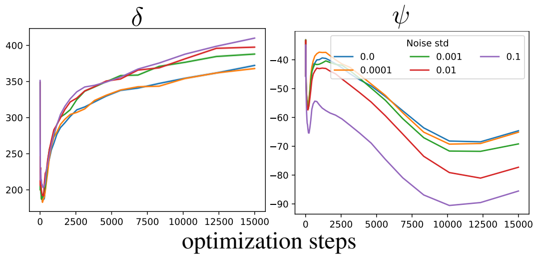

Figure 14: **Training dynamics of geometric descriptors for a VAE trained on MNIST with additive noise.** *As training progresses local complexity δ increases and local scaling ψ decreases suggesting an increase in expressivity and decrease in uncertainty on the data manifold.* At latter time-steps, $\psi \downarrow$ and $\delta \uparrow$ if noise std. is increased.

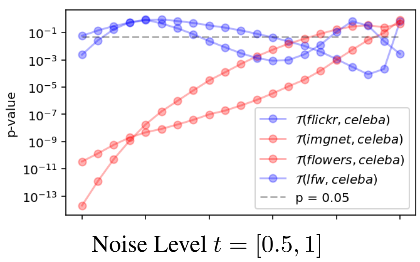

Figure 15: **Local scaling distribution difference between in-domain (blue) vs out-of-domain (red) datasets when conditioned on different noise levels** for an SD Unet trained on the CelebAHQ dataset. Here $\mathcal{T}(a, b)$ denotes a t-test between local scaling distributions for dataset a and dataset b.

Observations. By increasing the noise we control the puffiness of the target manifold. We observe that as the noise standard deviation is increased there is 1) increase in δ indicating the manifold becomes less smooth 2) decrease in local scaling indicating that the uncertainty decreases. We can also observe an initial dip in both local complexity and local scaling. This is similar to what was observed for discriminative models in (Humayun et al., 2024) where a double descent behavior was reported in the local complexity training dynamics of classification models. Based on these results, contrary to the observation in (Humayun et al., 2024), generative models do not have a double descent in local complexity however we do observe a double ascent in local scaling. *Our observations suggest that the training dynamics need to be taken into account, when comparing the local manifold geometry between two separately trained models.*

E Entropy difference between two nearby region

Suppose we have an injective $\mathcal{G} : Z \to X$ mapping learned by a CPWL generator $\mathcal{G}$. Any linear region ω in the latent space CPWL partition Ω is mapped to a unique region on the output manifold. We define S as:

$$S = \mathcal{G}(\mathbf{z}) \forall \mathbf{z} \in \omega = \mathbf{A}_\omega \mathbf{z} + \mathbf{b}_\omega \ \forall \mathbf{z} \in \omega$$

The change of volume from $\omega \to S$ is $\sqrt{det(\mathbf{A}_\omega{}^T \mathbf{A}_\omega)}$. Therefore for any latent z and output $x = \mathcal{G}(\mathbf{z})$:

$$p_{\mathcal{G}}(\mathbf{x}) = \sum_{\forall \omega \in \Omega} \frac{p_Z(\mathbf{z})}{\sqrt{det(\mathbf{A}_\omega{}^T \mathbf{A}_\omega)}} \mathbb{1}_{z \in \omega}$$

For any $\mathbf{z_1} \in \omega_1$ the sum from the above equation can be ignored, since for all other regions the value would be zero.

Taking negative log and expectation on both sides the conditional entropy becomes

$$H(p_{\mathcal{G}}(\mathbf{x_1}); \mathbf{z} \in \omega_1) = H(p_Z(\mathbf{z_1})) + log(\sqrt{det(\mathbf{A}_{\omega_1}{}^T \mathbf{A}_{\omega_1})})$$

For a uniform latent distribution and two regions ω_1 and ω_2, substituting the second term above with ψ_{ω_1}

$$H(p_{\mathcal{G}}(\mathbf{x_1}); \mathbf{z_1} \in \omega_1) - H(p_{\mathcal{G}}(\mathbf{x_2}); \mathbf{z_2} \in \omega_2) = \psi_{\omega_1} - \psi_{\omega_2}$$

F Broader Impact Statement

Our proposed framework for assessing and guiding generative models through manifold geometry offers several potential benefits to society. By providing a more objective and automated approach, we can significantly reduce the cost and time associated with human evaluation, making the auditing and mitigation of biases in large-scale models more accessible and efficient. This has implications for promoting fairness and equity in AI systems, particularly in domains where biases can have significant societal consequences.

Furthermore, our approach can empower researchers and practitioners to better understand the relationship between the geometry of learned representations and various aspects of model behavior, such as generation quality, diversity, and bias. This deeper understanding can inform the development of more robust and reliable generative models, leading to advancements in various fields, including art, design, healthcare, and education.

However, we recognize that our approach is not without limitations and potential risks. While it can be a valuable tool for identifying and mitigating biases, it should not and cannot fully replace human annotators, especially in high-risk domains where human judgment and contextual understanding are crucial. Our method focuses on reducing costs and improving the auditing process, but it should not be used as a standalone approach.

Moreover, the increased automation enabled by our approach raises concerns about the potential displacement of human annotators, leading to job losses and economic disruptions. While our method addresses some aspects of model evaluation, it is not comprehensive and cannot assess all facets of model behavior. Therefore, it should be used with caution and in conjunction with other evaluation methods, including human expertise.

G Extra Figures

Increasing ψ

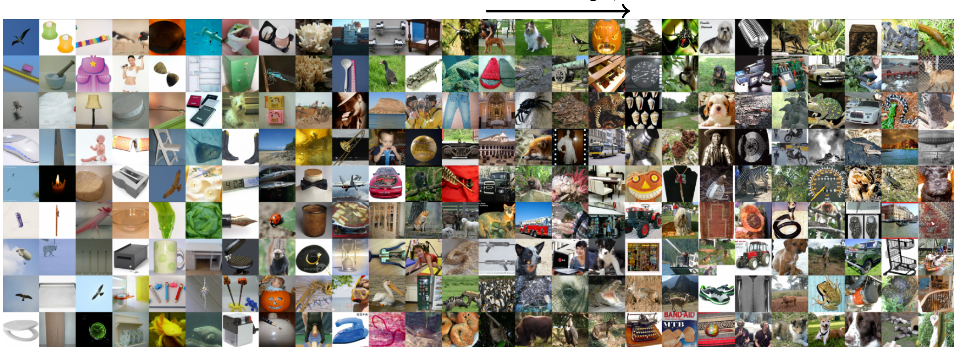

Figure 16: **Local Scaling is sensitive to natural image variations.** ImageNet images ordered along the columns (from left to right), with increasing local scaling ψ of the Stable Diffusion decoder learned manifold. We observe that ImageNet samples with lower values of ψ contain simpler backgrounds with modal representation of the object category. Conversely for higher ψ we have increasing diversity both in background and foreground features.

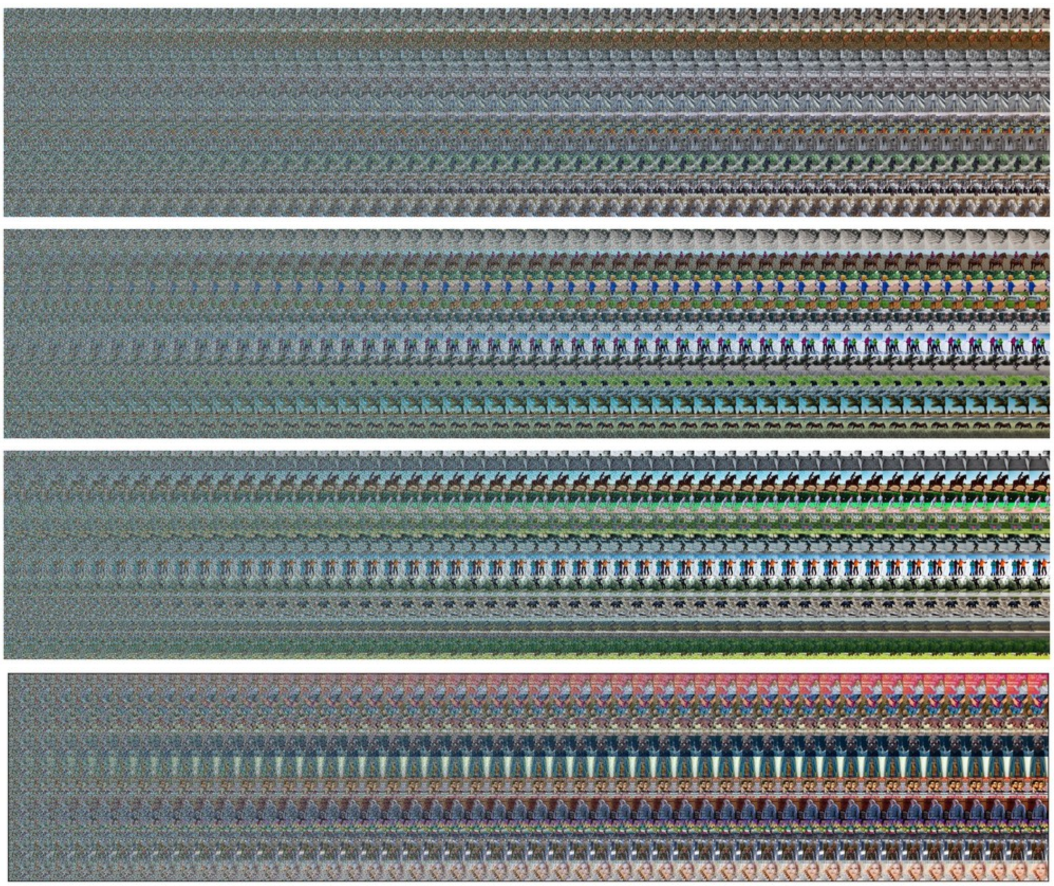

Figure 17: Images generated during 50 diffusion denoising steps for top to bottom, COCO prompts generated with guidance scale 1,5,9 and memorized prompts generated with guidance scale 7.5. Higher guidance scale images, as well as memorized images, tend to resolve faster during the denoising process.

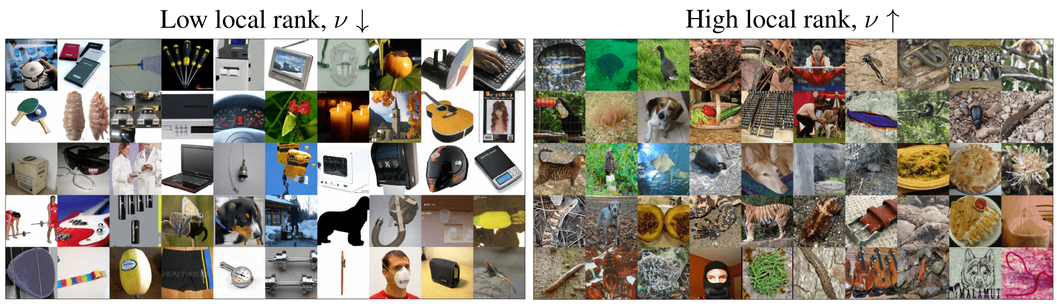

Figure 18: **Influence of the local rank descriptor value on overall image perception.** Images with the lowest (left) and highest (right) local rank ν from a set of 20000 randomly sampled ImageNet dataset samples. Low rank images contain simpler textures for every class compared to the high rank samples. This is because for images with higher local rank, the learned manifold is higher dimensional therefore allowing higher independent degrees of variations locally for the generated images.

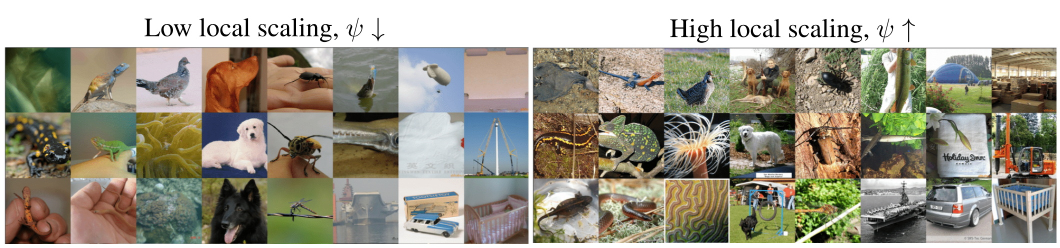

Figure 19: **Influence of the local scaling descriptor.**Imagenet images with high and low local scaling for the stable diffusion decoder. Each coordinate in both left and right image grids, correspond to the same imagenet class.

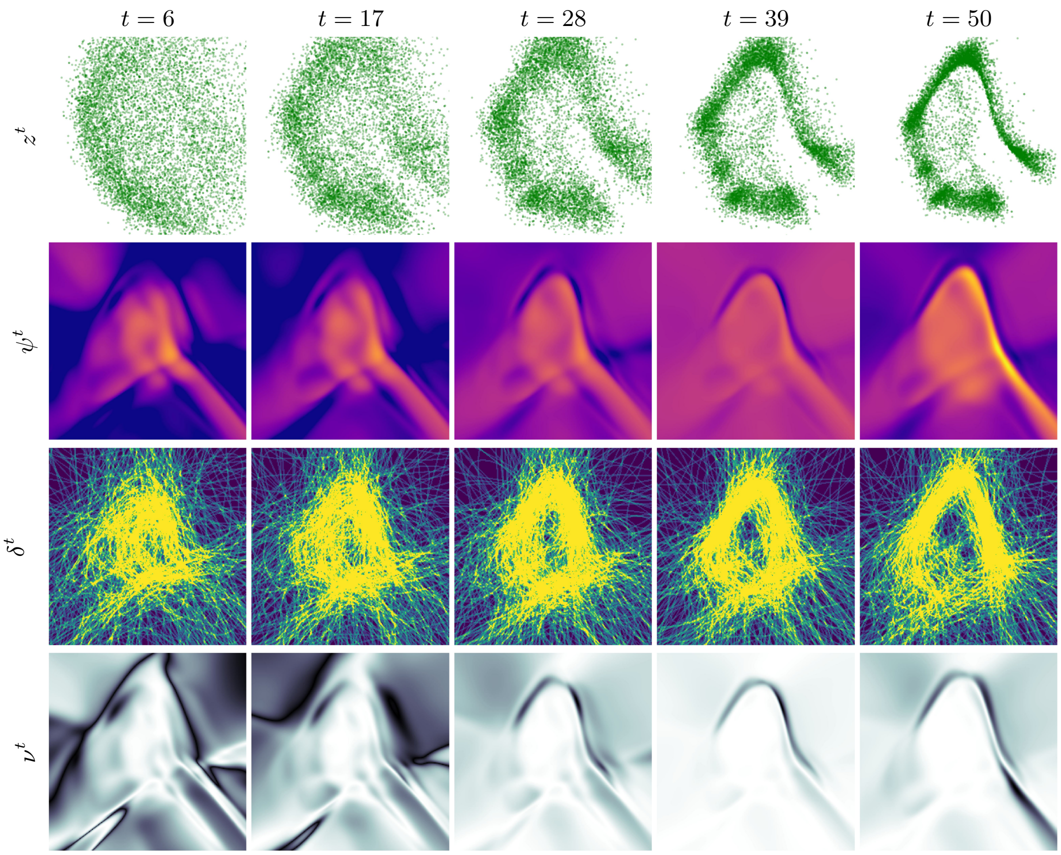

Figure 20: **2D Data Manifold Geometry, A toy Example.** After 11395 optimization steps. Geometry of a diffusion model input-output mapping, trained to on a toy 2D distribution. Local scaling lower around data manifold, local complexity higher around manifold, rank is lower around manifold as well. t=50 has considerably low variance in local scaling showing that final timestep has a diminishing change of density.

Figure 21: **Reward guidance on stable diffusion (maximizing the reward).**We observe a significant increase in both background detail and artifact diversity within the generated images.

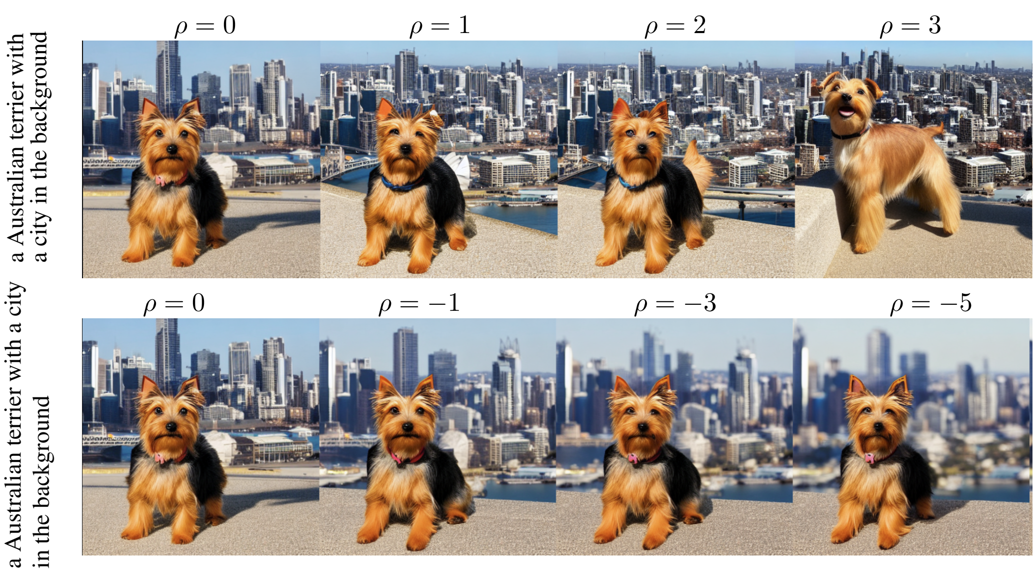

Figure 22: **Controlling mage diversity with local scaling.** Using Reward guidance to increase (top row) and decrease diversity (bottom row) using same initial seed.

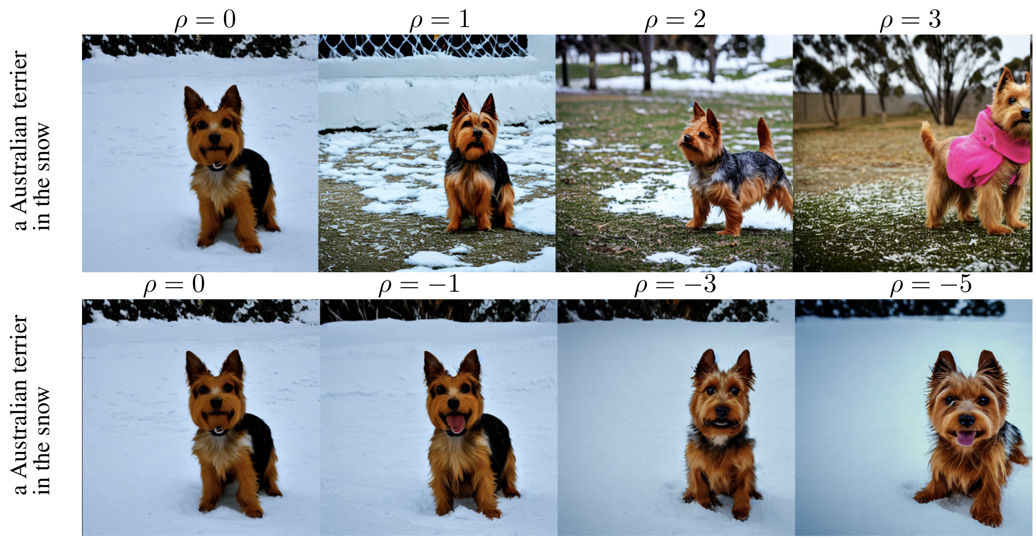

Figure 23: **Controlling mage diversity with local scaling.**Using Reward guidance to increase (top row) and decrease diversity (bottom row) using same initial seed.

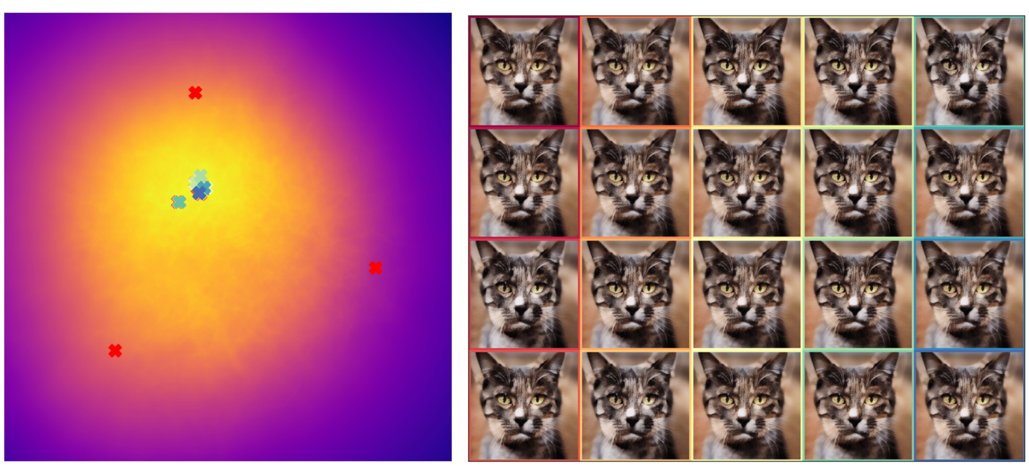

Figure 24: Decoded images (right) using 20 latents (left) from the 2D subspace, with highest ψ. Each image bounding box (right) is color coded according to the corresponding latent vector (left).

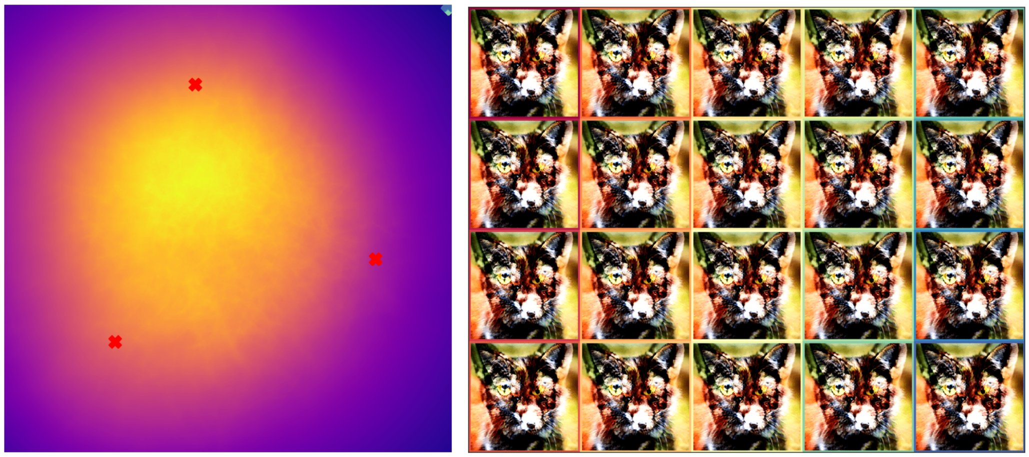

Figure 25: Decoded images (right) using 20 latents (left) from the 2D subspace, with lowest ψ. Each image bounding box (right) is color coded according to the corresponding latent vector (left). Selected latents lie outside the domain of the VQGAN latent space.

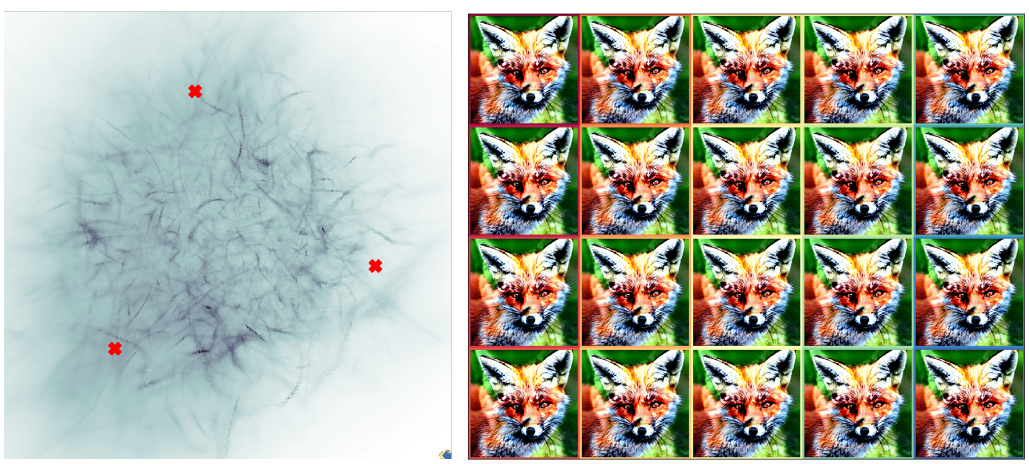

Figure 26: Decoded images (right) using 20 latents (left) from the 2D subspace, with highest ν. Each image bounding box (right) is color coded according to the corresponding latent vector (left). Selected latents lie outside the domain of the VQGAN latent space.

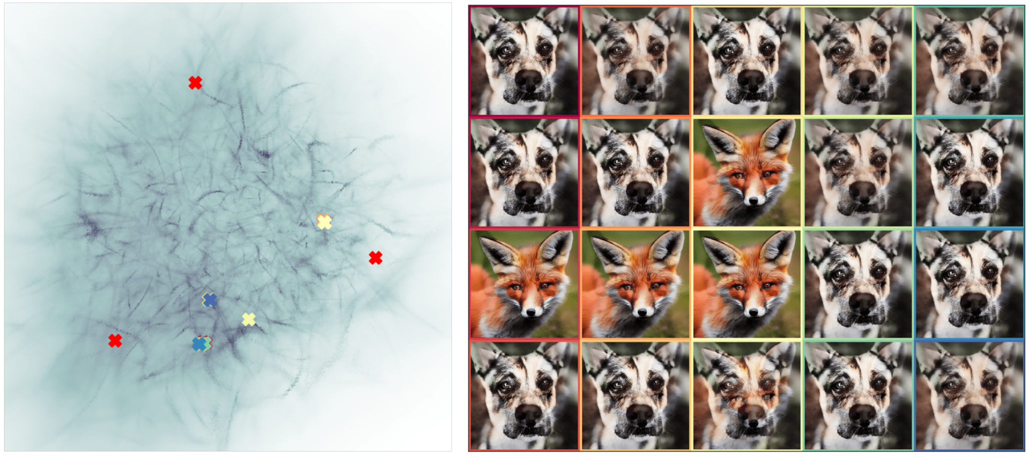

Figure 27: Decoded images (right) using 20 latents (left) from the 2D subspace, with lowest ν. Each image bounding box (right) is color coded according to the corresponding latent vector (left).

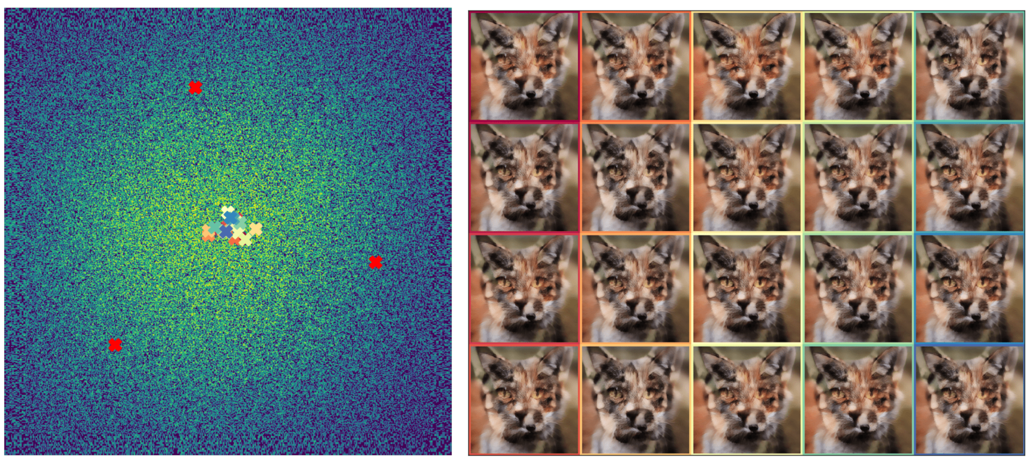

Figure 28: Decoded images (right) using 20 latents (left) from the 2D subspace, with highest δ. Each image bounding box (right) is color coded according to the corresponding latent vector (left).

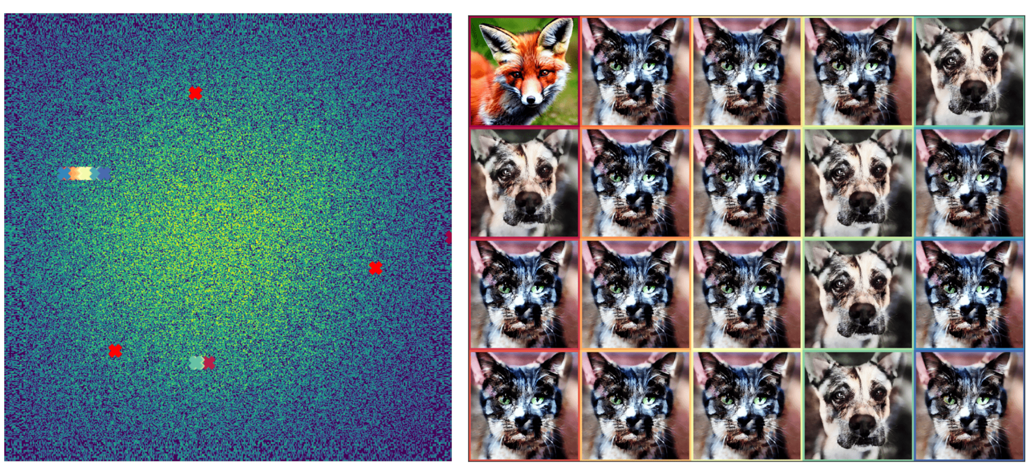

Figure 29: Decoded images (right) using 20 latents (left) from the 2D subspace, with lowest δ. Each image bounding box (right) is color coded according to the corresponding latent vector (left).

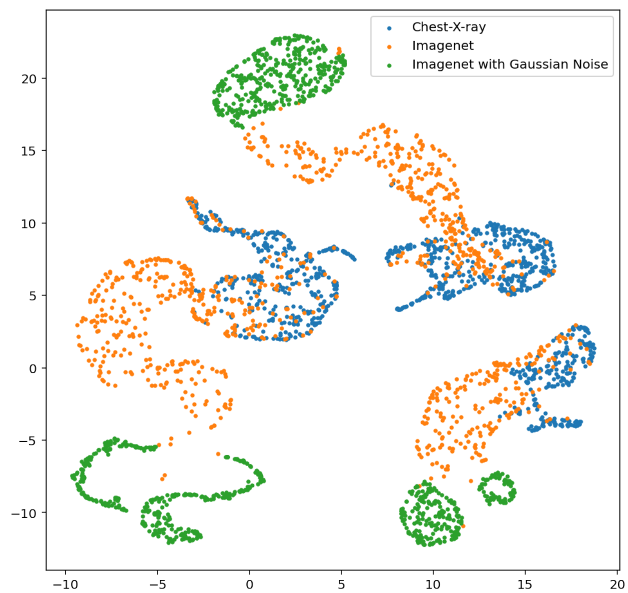

Figure 30: **UMAP visualization of the aggregated local geometry descriptors (local smoothness, local rank, and local complexity).** This reveals distinct, non-overlapping clusters, clearly separating the Imagenet, Imagenet Corrupted with Gaussian Noise, and Chest X-ray datasets. This visual evidence underscores the effectiveness of aggregating the descriptors to capture unique patterns within each dataset, demonstrating its ability to provide a meaningful and interpretable representation of the underlying data