IMAGE INTERPOLATION WITH SCORE-BASED RIEMANNIAN METRICS OF DIFFUSION MODELS

Shinnosuke Saito & Takashi Matsubara
Hokkaido University
{saitou.shinnosuke.y0@elms, matsubara@ist}.hokudai.ac.jp

ABSTRACT

Diffusion models excel in content generation by implicitly learning the data manifold, yet they lack a practical method to leverage this manifold—unlike other deep generative models equipped with latent spaces. This paper introduces a novel framework that treats the data space of pre-trained diffusion models as a Riemannian manifold, with a metric derived from the score function. Experiments with MNIST and Stable Diffusion show that this geometry-aware approach yields image interpolations that are more realistic, less noisy, and more faithful to prompts than existing methods, demonstrating its potential for improved content generation and editing.

1 INTRODUCTION

Deep generative models (DGMs) have achieved remarkable success in content generation across various domains (Rombach et al., 2022; Brooks et al., 2024; Tevet et al., 2023; Poole et al., 2023). They also offer applications such as image attribute editing (Kim et al., 2021), object replacement (Mokady et al., 2023), and smooth image interpolation (Zheng et al., 2024). This success can be explained through the lens of the manifold hypothesis, which states that high-dimensional data lie on lower-dimensional manifolds. DGMs equipped with latent spaces, including variational autoencoders (VAEs) (Kingma & Welling, 2014) and generative adversarial networks (GANs) (Goodfellow et al., 2014), learn to model such manifolds by embedding latent spaces into data spaces (Bengio et al., 2012; Dahal et al., 2022; Huang et al., 2022; Horvat & Pfister, 2022; Loaiza-Ganem et al., 2024). Given this ability, recent work has explored insights from differential geometry, e.g., introducing Riemannian metrics to the latent spaces of pre-trained DGMs and generating semantically consistent content through traversals based on geodesics (Shao et al., 2017; Chen et al., 2018; Arvanitidis et al., 2018; 2021; Fröhlich et al., 2021; Arvanitidis et al., 2022; Lee et al., 2022).

Diffusion models (Sohl-Dickstein et al., 2015; Ho et al., 2020; Song et al., 2021a;b), a class of DGMs known for state-of-the-art generation quality, are also considered to learn data manifolds. Some studies estimated their intrinsic dimensionality (Stanczuk et al., 2024; Kamkari et al., 2024; Horvat & Pfister, 2024; Ventura et al., 2025), and others improved the image quality by projecting samples onto the data manifolds during the generation (Chung et al., 2022; He et al., 2024; Zirvi et al., 2025). Nonetheless, they have not fully exploited the underlying Riemannian structures for tasks beyond naive sampling (Park et al., 2023a).

In this paper, we introduce a Riemannian metric on the data space by leveraging the score function of diffusion models. We examined a small model trained on MNIST (Deng, 2012) and Stable Diffusion (Rombach et al., 2022) to demonstrate that our method yields more natural and faithful transitions, as assessed with CLIP-IQA (Wang et al., 2023), compared with existing methods: linear (Lerp) (Ho et al., 2020) and spherical linear (Slerp) interpolation (Song et al., 2021a), as well as NAO (Samuel et al., 2023) and Noise Diffusion (Zheng et al., 2024). We also find that NAO and NoiseDiffusion suffer from severe reconstruction errors, whereas the other methods do not, as evaluated by mean squared error (MSE), LPIPS (Zhang et al., 2018), and DreamSim (Fu et al., 2023).

2 RELATED WORK

Deep Generative Models through Riemannian Geometry Pre-trained VAEs and GANs are known to be capable of editing generated image attributes (e.g., facial expressions and hair color) by linearly manipulating latent variables (Härkönen et al., 2020; Voynov & Babenko, 2020; Zhu et al., 2022; Shen & Zhou, 2021; Zhu et al., 2021). To mitigate the constraint of being linear (Arvanitidis et al., 2018), some studies treated the latent space as a learned manifold with a Riemannian metric, improving the edit quality (Arvanitidis et al., 2021; Fröhlich et al., 2021; Arvanitidis et al., 2022). A typical way defines a metric by pulling back the Euclidean metric from the data space, but it still assumed a linear data space, not fully capturing the underlying geometric structure (Shao et al., 2017; Chen et al., 2018; Arvanitidis et al., 2018). Another way learns a metric on the data space and then pullback it to the latent space (Arvanitidis et al., 2021), but it requires task-specific metric learning and additional architectures. In diffusion models, the bottleneck layer of the U-Net for noise-prediction are suggested to serve as a latent space (Kwon et al., 2023). Some studies assume a Euclidean metric on the bottleneck layer and pullback it to the data space, defining a metric on the data space (Park et al., 2023a;b). Like the traditional ways for VAEs and GANs, this approach is however limited by the assumption of a linear latent space.

Data Manifolds of Diffusion Models Diffusion models are considered to learn data distributions $p_t(x_t)$ and the underlying manifolds at each diffusion time step t (Pidstrigach, 2022; Loaiza-Ganem et al., 2024). Several studies have estimated the local dimensionality of a data manifold $\mathcal{M}_t$ in diffusion models by examining the trained score function $s_\theta(x_t, t) \approx \nabla_x \log p_t(x_t, t)$, specifically its Jacobian $\nabla_x s_\theta(x_t, t)$ (essentially the Hessian of $\log p_t(x_t, t)$) (Stanczuk et al., 2024; Ventura et al., 2025) or its divergence (Kamkari et al., 2024). Other works suggest that lower-quality samples arise when the reverse process drifts away from the manifold. To mitigate this, some studies project the generated image x_t onto the manifold $\mathcal{M}_t$ by assuming a linear manifold (Chung et al., 2022), by using a separate autoencoder (He et al., 2024), or by constructing a subspace via singular value decomposition (Zirvi et al., 2025).

Beyond naive sampling, some studies have tackled image interpolation (Deschenaux et al., 2024). Several methods require retraining (Zhang et al., 2023; Yang et al., 2024), depend on specific models (Preechakul et al., 2022; Kim et al., 2025; Lu et al., 2024), or need additional conditioning (Wang & Golland, 2023), and hence do not fully leverage the intrinsic manifold structure of pre-trained diffusion models. The simplest method, Lerp (Ho et al., 2020), linearly interpolates two images x_t, x'_t after t diffusion steps and then applies the reverse process. This effectively treats the data space at time t (often called a noise space) as a linear latent space, similar to VAEs or GANs. Because the norm of a latent variable correlates with semantic richness (Samuel et al., 2023; Alper & Averbuch-Elor, 2024), Lerp leads to blur or loss of detail by decreasing norms. Slerp (Shoemake, 1985) preserves norms by interpolating along a spherical path. NAO leverages the fact that norms of samples drawn from a normal distribution follow a chi distribution and maximizes the probability of the path between two endpoints (Samuel et al., 2023). While these methods often yield smoother transitions than Lerp, they still lose detail and further generate artifacts because the latent variables for natural images often deviate from the expected normal distribution. NoiseDiffusion addressed this by adding extra noise and clipping extreme noise (Zheng et al., 2024), which indeed generates high-quality images but not necessarily a proper interpolation, as it can inject or remove information. Ultimately, no existing interpolation method fully exploits the manifold structure in the data space.

3 METHOD: GEODESIC INTERPOLATION

Riemannian Metric based on Score Function For completeness, we provide background details on diffusion models and Riemannian geometry in Appendix A. Here, we focus on our main proposal. We propose the metric tensor g in the data space $\mathcal{M}_t$ at time t of a diffusion model, which is represented by a matrix

$$G_{x_t} = J_{x_t}^\top J_{x_t} \quad \text{for} \quad J_{x_t} = \nabla_{x_t} s_\theta(x_t, t). \tag{1}$$

Since the score function $s_\theta(x_t, t)$ is an approximation of $\nabla_{x_t} \log p_t(x_t)$, its Jacobian J_{x_t} corresponds to the Hessian $H_{x_t} = \nabla_{x_t} \nabla_{x_t} \log p_t(x_t)$. As long as J_{x_t} is non-degenerate, G_{x_t} is positive definite and thus valid as a Riemannian metric. With this metric, the length of a vector v at x_t is

$|v|_g = \sqrt{\langle v, v\rangle_g} = \|J_{x_t} v\|_2$. The length of a curve is obtained by integrating local lengths along the curve. The length-minimizing curve between two points is called a *geodesic*. Thus, we define an interpolation between two samples $x_t^{(0)}$ and $x_t^{(N)}$ as the geodesic under the metric g given by Eq. (1).

Interpretation of Proposed Metric The proposed metric is apparently similar to the pullback metric in prior works (Shao et al., 2017; Chen et al., 2018; Arvanitidis et al., 2018; 2021; Park et al., 2023a;b). However, these works use the Jacobian of a map from the data space to a latent space (or vice versa), whereas our metric employs the Jacobian of the score function and never pullbacks any predefined metric. Note also that it is not the Hessian metric for a Hessian manifold.

In reality, observational noise prevents the data distribution from forming a perfectly low-dimensional manifold; the distribution is effectively collapsed (or compressed) along certain directions. Since $s_\theta(x_t, t)$ is the gradient of the log-likelihood $\log p_t(x_t)$, it points in such directions to the manifold, guiding samples along directions of higher density. Consequently, the directions corresponding to large eigenvalues of J_{x_t} indicate collapsed dimensions, whereas those corresponding to small eigenvalues are tangential to the manifold (Stanczuk et al., 2024; Ventura et al., 2025). Therefore, following directions for which $\|J_{x_t} v\|$ is small can thus be seen as moving *within* or *parallel to* the manifold, providing a smooth transition of images.

Another interpretation follows from the Taylor expansion of s_θ around x, which yields $\|s_\theta(x + v, t) - s_\theta(x, t)\|_2 = \|J_x v\|_2 + O(\|v\|_2^2)$. This implies that the proposed geodesic corresponds to a curve along which $s_\theta(x, t)$ changes as little as possible. Earlier studies have shown that the gradient of a log-likelihood (with respect to model parameters) can serve as a robust, semantically meaningful representation of input (Charpiat et al., 2019; Hanawa et al., 2021; Yeh et al., 2018). In this light, our metric can be viewed as a measure of the *semantic closeness* between infinitesimally different samples, providing transitions that preserve the underlying meaning within the data manifold.

Implementation Let $s \in [0, 1]$ be the independent variable parameterizing a curve $\gamma : s(\in [0, 1]) \mapsto \gamma(s)$. The curve γ is discretized as a sequence of data points $x_t^{(0)}, \ldots, x_t^{(N)}$, where $\gamma(s_i) = x^{(i)}$, $s_0 = 0$, $s_N = 1$, and $s_{i+1} - s_i = \Delta s$. The length of the curve, $L[\gamma]$, is numerically approximated using the trapezoidal rule:

$$L[\gamma] = \int_0^1 l(s)\mathrm{d}s \approx \sum_{i=0}^{N-1} \tfrac{1}{2}(l(s_{i+1}) + l(s_i))\Delta s, \tag{2}$$

where the local path length $l(s_i)$ is given by $l(s_i) = \sqrt{\gamma'(s_i)^\top G_{\gamma(s_i)} \gamma'(s_i)} \approx \sqrt{(v_t^{(i)})^\top G_{x_t^{(i)}} v_t^{(i)}}$, and $v_t^{(i)}$ denotes the velocity at point $x_t^{(i)}$. This is easily computed by the Jacobian-vector product. The velocities $v_t^{(i)}$ are approximated using second-order finite differences.

Given two samples, $x_t^{(0)}$ and $x_t^{(N)}$, the geodesic path is obtained by minimizing the discrete approximation of $L[\gamma]$ with respect to the intermediate points $x_t^{(1)}, \ldots, x_t^{(N-1)}$, i.e.,

$$\min_{x_t^{(1)}, \ldots, x_t^{(N-1)}} L[\gamma] \quad \text{s.t. } \gamma(s_0) = x_t^{(0)}, \gamma(s_N) = x_t^{(N)}. \tag{3}$$

To prevent the intermediate points from collapsing to a single point, we add the variance of the Euclidean distance, $\mathrm{Var}\left[\|x_t^{(i+1)} - x_t^{(i)}\|_2\right]$, as a regularization term to the loss function, multiplied by hyperparameter λ. This term guides the velocity to be constant. The initial values of $x_t^{(1)}, \ldots, x_t^{(N-1)}$ can be initialized using any reasonable method; in this work, we employed Slerp (Song et al., 2021a).

We observed that directly computing the geodesic in the raw data space at $t = 0$ leads to poor results due to the highly rugged landscape of the score function s_θ at $t = 0$, likely because it memorizes individual training data points. To address this issue, we employed DDIM inversion to map samples into the data space at a specific time step $t = \tau > 0$, compute the geodesic there, and then apply the reverse DDIM process to obtain the final image sequence in the original data space at $t = 0$.

4 Experiments and Results

We evaluated our method with Stable Diffusion (Rombach et al., 2022), as well as Lerp (Ho et al., 2020), Slerp (Shoemake, 1985), NAO (Samuel et al., 2023), and NoiseDiffusion (Zheng et al., 2024). We set max forward steps to $T = 50$ and the number of forward steps before interpolation to $\tau = 50$ for NAO and 30 for others. We set the number of interpolation steps to $N = 10$. We optimized

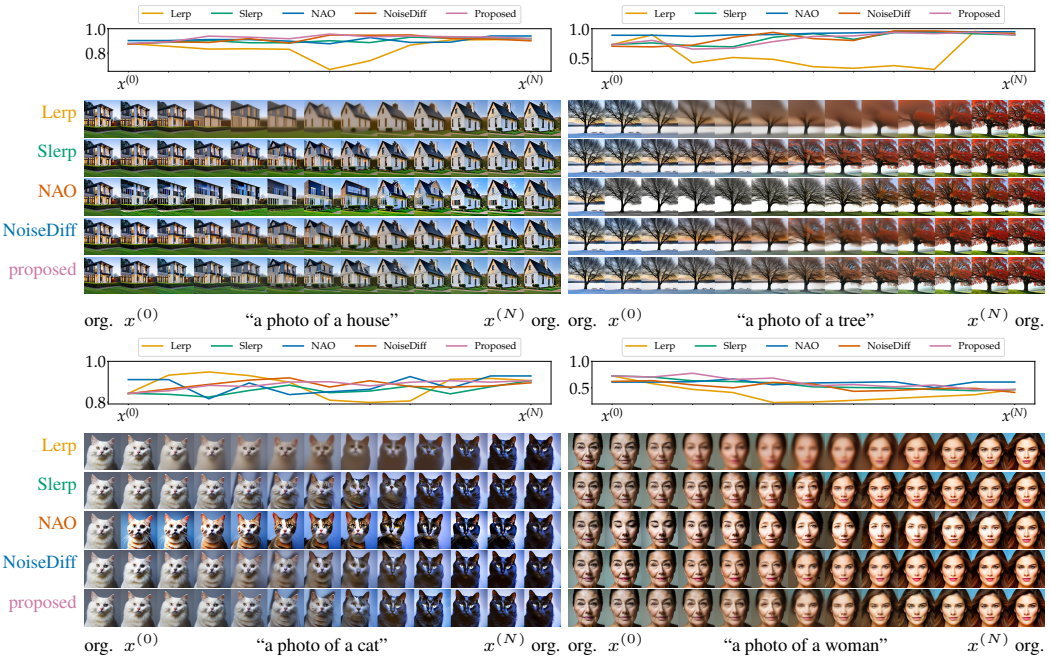

Figure 1: Images generated by Stable Diffusion with prompts shown below, with interpolations using different methods (CFG: 7.5). Images at both ends are original, adjacent to them are the reconstructions, and in between are the interpolation results. The sample-wise fidelity is visualized above each image as a graph.

Table 1: Mean Reconstruction Errors and CLIP-IQA Assessment for Four Examples in Figure 1

	Reconstruction Errors			CLIP-IQA		
Method	**MSE** $[\times 10^{-3}]$	**LPIPS** $[\times 10^{-1}]$	**DreamSim** $[\times 10^{-2}]$	**Reality**	**Noisiness**	**Fidelity**
Lerp	**7.89**	**1.71**	**5.34**	0.389	0.406	0.686
Slerp	**7.89**	**1.71**	**5.34**	0.704	0.765	0.784
NAO	64.04	6.03	32.82	0.617	0.766	**0.815**
NoiseDiff	14.13	2.35	6.93	0.600	0.646	0.783
Proposed	**7.89**	**1.71**	**5.34**	**0.716**	**0.818**	0.810

the path using Adam (Kingma & Ba, 2015) for 5,000 iterations. The learning rate was initialized at 10^{-2} and decayed to zero using cosine annealing (Loshchilov & Hutter, 2017).

Figure 1 and Table 1 summarize the results. As previous works have shown, Lerp produces blurry images with a notable loss of detail. Slerp yields smoother transitions, but sometimes objects appear doubled, like houses or trees. NAO and NoiseDiffusion suffer from severe reconstruction errors because of a long diffusion process and added or clipped noise, as evaluated by MSE, LPIPS (Zhang et al., 2018), and DreamSim (Fu et al., 2023). In contrast, the proposed method demonstrates the most realistic and faithful transitions, gradually adjusting objects, color, and lighting while avoiding noise, as assessed with CLIP-IQA. See Appendix C for additional explanations and results.

5 Conclusion

This paper introduces a Riemannian metric derived from the score function of pre-trained diffusion models. The metric yields geodesic paths that naturally follow the learned data manifold, providing a geometry-aware framework for image interpolation. Experiments on MNIST and Stable Diffusion show smooth, natural, and faithful transitions than existing methods. The proposed geometric framework has broader potential applications, such as video editing by treating a video as a curve on the manifold and using parallel transport to modify all frames simultaneously.

REFERENCES

Morris Alper and Hadar Averbuch-Elor. Emergent Visual-Semantic Hierarchies in Image-Text Representations. In *European Conference on Computer Vision (ECCV)*, 2024.

Georgios Arvanitidis, Lars Kai Hansen, and Søren Hauberg. Latent Space Oddity: on the Curvature of Deep Generative Models. In *International Conference on Learning Representations (ICLR)*, 2018.

Georgios Arvanitidis, Soren Hauberg, and Bernhard Schölkopf. Geometrically Enriched Latent Spaces. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2021.

Georgios Arvanitidis, Bogdan M. Georgiev, and Bernhard Schölkopf. A Prior-Based Approximate Latent Riemannian Metric. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2022.

Yoshua Bengio, Aaron C. Courville, and Pascal Vincent. Representation Learning: A Review and New Perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 35: 1798–1828, 2012.

Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video Generation Models as World Simulators. 2024.

Guillaume Charpiat, Nicolas Girard, Loris Felardos, and Yuliya Tarabalka. Input Similarity from the Neural Network Perspective. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Nutan Chen, Alexej Klushyn, Richard Kurle, Xueyan Jiang, Justin Bayer, and Patrick Smagt. Metrics for Deep Generative Models. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2018.

Hyungjin Chung, Byeongsu Sim, Dohoon Ryu, and Jong Chul Ye. Improving Diffusion Models for Inverse Problems using Manifold Constraints. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Biraj Dahal, Alexander Havrilla, Minshuo Chen, Tuo Zhao, and Wenjing Liao. On Deep Generative Models for Approximation and Estimation of Distributions on Manifolds. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Li Deng. The Mnist Database of Handwritten Digit Images for Machine Learning Research. *IEEE Signal Processing Magazine*, 29:141–142, 2012.

Justin Deschenaux, Igor Krawczuk, Grigorios G. Chrysos, and Volkan Cevher. Going beyond Compositions, DDPMs Can Produce Zero-Shot Interpolations. In *International Conference on Machine Learning (ICML)*, 2024.

Christian Fröhlich, Alexandra Gessner, Philipp Hennig, Bernhard Schölkopf, and Georgios Arvanitidis. Bayesian Quadrature on Riemannian Data Manifolds. In *International Conference on Machine Learning (ICML)*, 2021.

Stephanie Fu, Netanel Tamir, Shobhita Sundaram, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola. Dreamsim: Learning New Dimensions of Human Visual Similarity using Synthetic Data. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative Adversarial Nets. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2014.

Kazuaki Hanawa, Sho Yokoi, Satoshi Hara, and Kentaro Inui. Evaluation of Similarity-based Explanations. In *International Conference on Learning Representations (ICLR)*, 2021.

Erik Härkönen, Aaron Hertzmann, Jaakko Lehtinen, and Sylvain Paris. GANSpace: Discovering Interpretable GAN Controls. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Yutong He, Naoki Murata, Chieh-Hsin Lai, Yuhta Takida, Toshimitsu Uesaka, Dongjun Kim, Wei-Hsiang Liao, Yuki Mitsufuji, J Zico Kolter, Ruslan Salakhutdinov, and Stefano Ermon. Manifold Preserving Guided Diffusion. In *International Conference on Learning Representations (ICLR)*, 2024.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising Diffusion Probabilistic Models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Christian Horvat and Jean-Pascal Pfister. Density Estimation on Low-dimensional Manifolds: An Inflation-Deflation Approach. *Journal of Machine Learning Research (JMLR)*, 2022.

Christian Horvat and Jean-Pascal Pfister. On Gauge Freedom, Conservativity and Intrinsic Dimensionality Estimation in Diffusion models. In *International Conference on Learning Representations (ICLR)*, 2024.

Jian Huang, Yuling Jiao, Zhen Li, Shiao Liu, Yang Wang, and Yunfei Yang. An Error Analysis of Generative Adversarial Networks for Learning Distributions. *Journal of Machine Learning Research (JMLR)*, 2022.

Hamidreza Kamkari, Brendan Leigh Ross, Rasa Hosseinzadeh, Jesse C. Cresswell, and Gabriel Loaiza-Ganem. A Geometric View of Data Complexity: Efficient Local Intrinsic Dimension Estimation with Diffusion Models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Gwanghyun Kim, Taesung Kwon, and Jong-Chul Ye. DiffusionCLIP: Text-Guided Diffusion Models for Robust Image Manipulation. *the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2416–2425, 2021.

Yeongmin Kim, Kwanghyeon Lee, Minsang Park, Byeonghu Na, and Il chul Moon. Diffusion Bridge AutoEncoders for Unsupervised Representation Learning. In *International Conference on Learning Representations (ICLR)*, 2025.

Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. In *International Conference on Learning Representations (ICLR)*, 2015.

Diederik P. Kingma and Max Welling. Auto-Encoding Variational Bayes. In *International Conference on Learning Representations (ICLR)*, 2014.

Mingi Kwon, Jaeseok Jeong, and Youngjung Uh. Diffusion Models Already Have A Semantic Latent Space. In *International Conference on Learning Representations (ICLR)*, 2023.

John M. Lee. *Introduction to Riemannian Manifolds*. Springer, 2019.

Yonghyeon Lee, Seungyeon Kim, Jinwon Choi, and Frank Park. A Statistical Manifold Framework for Point Cloud Data. In *International Conference on Machine Learning (ICML)*, 2022.

Gabriel Loaiza-Ganem, Brendan Leigh Ross, Rasa Hosseinzadeh, Anthony L. Caterini, and Jesse C. Cresswell. Deep Generative Models through the Lens of the Manifold Hypothesis: A Survey and New Connections. *Transactions on Machine Learning Research (TMLR)*, 2024.

Ilya Loshchilov and Frank Hutter. SGDR: Stochastic Gradient Descent with Warm Restarts. In *International Conference on Learning Representations (ICLR)*, 2017.

Zeyu Lu, Chengyue Wu, Xinyuan Chen, Yaohui Wang, Lei Bai, Yu Qiao, and Xihui Liu. Hierarchical Diffusion Autoencoders and Disentangled Image Manipulation. In *IEEE Workshop/Winter Conference on Applications of Computer Vision (WACV)*, pp. 5362–5371, 2024.

Calvin Luo. Understanding Diffusion Models: A Unified Perspective. *arXiv*, 2022.

Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-Text Inversion for Editing Real Images using Guided Diffusion Models. In *the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6038–6047, 2023.

Yong-Hyun Park, Mingi Kwon, Jaewoong Choi, Junghyo Jo, and Youngjung Uh. Understanding the Latent Space of Diffusion Models through the Lens of Riemannian Geometry. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023a.

Yong-Hyun Park, Mingi Kwon, Junghyo Jo, and Youngjung Uh. Unsupervised Discovery of Semantic Latent Directions in Diffusion Models, 2023b.

Jakiw Pidstrigach. Score-Based Generative Models Detect Manifolds. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. DreamFusion: Text-to-3D using 2D Diffusion. In *International Conference on Learning Representations (ICLR)*, 2023.

Konpat Preechakul, Nattanat Chatthee, Suttisak Wizadwongsa, and Supasorn Suwajanakorn. Diffusion Autoencoders: Toward a Meaningful and Decodable Representation. In *the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision. In *International Conference on Machine Learning (ICML)*, 2021.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution Image Synthesis with Latent Diffusion Models. In *the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Dvir Samuel, Rami Ben-Ari, Nir Darshan, Haggai Maron, and Gal Chechik. Norm-guided Latent Space Exploration for Text-to-image Generation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Hang Shao, Abhishek Kumar, and P. Thomas Fletcher. The Riemannian Geometry of Deep Generative Models. *the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2017.

Yujun Shen and Bolei Zhou. Closed-form Factorization of Latent Semantics in GANs. In *the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

Ken Shoemake. Animating Rotation with Quaternion Curves. *Conference on Computer Graphics and Interactive Techniques (SIGGRAPH)*, 1985.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep Unsupervised Learning using Nonequilibrium Thermodynamics. In *International Conference on Learning Representations (ICLR)*, 2015.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising Diffusion Implicit Models. In *International Conference on Learning Representations (ICLR)*, 2021a.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-Based Generative Modeling through Stochastic Differential Equations. In *International Conference on Learning Representations (ICLR)*, 2021b.

Jan Pawel Stanczuk, Georgios Batzolis, Teo Deveney, and Carola-Bibiane Schönlieb. Diffusion Models Encode the Intrinsic Dimension of Data Manifolds. In *International Conference on Machine Learning (ICML)*, 2024.

Guy Tevet, Sigal Raab, Brian Gordon, Yoni Shafir, Daniel Cohen-or, and Amit Haim Bermano. Human Motion Diffusion Model. In *International Conference on Learning Representations (ICLR)*, 2023.

Enrico Ventura, Beatrice Achilli, Gianluigi Silvestri, Carlo Lucibello, and Luca Ambrogioni. Manifolds, Random Matrices and Spectral Gaps: The Geometric Phases of Generative Diffusion. In *International Conference on Learning Representations (ICLR)*, 2025.

Andrey Voynov and Artem Babenko. Unsupervised Discovery of Interpretable Directions in the GAN Latent Space. In *International Conference on Machine Learning (ICML)*, 2020.

Clinton J. Wang and Polina Golland. Interpolating between Images with Diffusion Models. In *ICML 2023 Workshop on Challenges of Deploying Generative AI*, 2023.

Jianyi Wang, Kelvin C.K. Chan, and Chen Change Loy. Exploring CLIP for Assessing the Look and Feel of Images. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2023.

Zhaoyuan Yang, Zhengyang Yu, Zhiwei Xu, Jaskirat Singh, Jing Zhang, Dylan Campbell, Peter Tu, and Richard Hartley. IMPUS: Image Morphing with Perceptually-Uniform Sampling Using Diffusion Models. In *International Conference on Learning Representations (ICLR)*, 2024.

Chih-Kuan Yeh, Joon Sik Kim, Ian E. H. Yen, and Pradeep Ravikumar. Representer Point Selection for Explaining Deep Neural Networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Kaiwen Zhang, Yifan Zhou, Xudong Xu, Xingang Pan, and Bo Dai. DiffMorpher: Unleashing the Capability of Diffusion Models for Image Morphing. In *the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

PengFei Zheng, Yonggang Zhang, Zhen Fang, Tongliang Liu, Defu Lian, and Bo Han. NoiseDiffusion: Correcting Noise for Image Interpolation with Diffusion Models beyond Spherical Linear Interpolation. In *International Conference on Learning Representations (ICLR)*, 2024.

Jiapeng Zhu, Ruili Feng, Yujun Shen, Deli Zhao, Zheng-Jun Zha, Jingren Zhou, and Qifeng Chen. Low-rank Subspaces in GANs. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pp. 16648–16658, 2021.

Jiapeng Zhu, Yujun Shen, Yinghao Xu, Deli Zhao, and Qifeng Chen. Region-Based Semantic Factorization in GANs. In *International Conference on Machine Learning (ICML)*, volume 162, pp. 27612–27632, 2022.

Rayhan Zirvi, Bahareh Tolooshams, and Anima Anandkumar. Diffusion State-Guided Projected Gradient for Inverse Problems. In *International Conference on Learning Representations (ICLR)*, 2025.

A BACKGROUND THEORY

This section provides the foundational concepts necessary for our proposed method. We begin with an overview of diffusion models, describing their forward and reverse processes and the link between noise prediction and the score function. We then introduce key elements of Riemannian geometry, focusing on how Riemannian metrics induce distances and paths on manifolds.

A.1 DIFFUSION MODELS

Diffusion models are a class of DGMs inspired by non-equilibrium thermodynamics (Sohl-Dickstein et al., 2015; Ho et al., 2020). The model consists of two processes: a forward process that adds noise to the data and a reverse process that removes the noise. Their core idea is to model the underlying data distribution by denoising noisy samples.

Forward Process Let $x_0 \in \mathbb{R}^D$ be a data sample. The forward process is defined as a Markov chain in which Gaussian noise is added at each time step $t = 1, \ldots, T$:

$$q(x_t|x_{t-1}) = \mathcal{N}\left(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t I\right) = \mathcal{N}\left(\sqrt{\frac{\alpha_t}{\alpha_{t-1}}}x_{t-1}, \left(1 - \frac{\alpha_t}{\alpha_{t-1}}\right) I\right), \tag{4}$$

where $\{\beta_t\}_{t=1}^T$ is a variance schedule, I is the identity matrix in $\mathbb{R}^D$, and $\alpha_t = \prod_{s=1}^t (1-\beta_s)$. As t increases, x_t becomes progressively more corrupted by noise until x_T is nearly an isotropic Gaussian distribution.

Reverse Process To invert this forward process, a reverse Markov chain $p_\theta(x_{t-1}|x_t)$ from $x_T \sim \mathcal{N}(0, I)$ is constructed as

$$x_{t-1} = \frac{1}{\sqrt{1-\beta_t}}\left(x_t - \frac{\beta_t}{\sqrt{1-\alpha_t}}\epsilon_\theta(x_t, t)\right) + \sigma_t z_t, \tag{5}$$

with trainable noise predictor ϵ_θ, where $z_t \sim \mathcal{N}(0, I)$ and $\sigma_t^2 = \beta_t$ is a variance. The noise predictor $\epsilon_\theta(x_t, t)$ is trained by minimizing the objective:

$$\mathcal{L}(\theta) = \mathbb{E}_{x,\epsilon_t,t}\left[\|\epsilon_t - \epsilon_\theta(x_t, t)\|_2^2\right], \tag{6}$$

where $\epsilon_t \sim \mathcal{N}(0, I)$ is the noise added during the forward process at time step t.

Denoising Diffusion Implicit Models Denoising diffusion implicit models (DDIMs) (Song et al., 2021a) modified Eq. (4) as a non-Markovian process $q(x_{t-1}|x_t, x_0) = \mathcal{N}(\sqrt{\alpha_{t-1}}x_0 + \sqrt{1-\alpha_{t-1}-\sigma_t^2}, \sigma_t^2 I)$. Then the reverse process becomes

$$x_{t-1} = \sqrt{\alpha_{t-1}}\left(\frac{x_t - \sqrt{1-\alpha_t}\epsilon_\theta(x_t, t)}{\sqrt{\alpha_t}}\right) + \sqrt{1-\alpha_{t-1}-\sigma_t^2} \cdot \epsilon_\theta(x_t, t) + \sigma_t z_t, \tag{7}$$

where $\sigma_t = \eta\sqrt{(1-\alpha_{t-1})/(1-\alpha_t)}\sqrt{1-\alpha_t/\alpha_{t-1}}$. Here, $\eta \in [0, 1]$ determines the stochasticity: $\eta = 1$ recovers the DDPM, while $\eta = 0$ yields a deterministic update.

Formulation as Stochastic Differential Equations Diffusion models can also be formulated using stochastic differential equations (SDEs) (Song et al., 2021b). In that viewpoint, the forward process is governed by a continuous-time SDE, and its time-reversal is defined through the corresponding reverse-time SDE, which depends on the score function $\nabla_{x_t} \log p_t(x_t)$, where $p_t(\cdot)$ denotes the distribution of x_t at time t. Notably, the noise-prediction network ϵ_θ is closely tied to the score function (Luo, 2022) as:

$$\nabla_{x_t} \log p_t(x_t) \approx -\frac{1}{\sqrt{1-\alpha_t}}\epsilon_\theta(x_t, t). \tag{8}$$

Hence, learning ϵ_θ for noise prediction is equivalent to learning the score function.

Conditional Generation

$$\tilde{\epsilon}_t(x_t, t, C) = \epsilon_t(x_t, t, \emptyset) + \gamma \cdot (\epsilon_t(x_t, t, C) - \epsilon_t(x_t, t, \emptyset)) \tag{9}$$

A.2 RIEMANNIAN GEOMETRY

We adopt the presentation in Lee. (2019). Let $\mathcal{M}$ be a smooth manifold. A *Riemannian metric* g on $\mathcal{M}$ is a smooth $(0,2)$-tensor field such that at every point $p \in \mathcal{M}$, the tensor g_p defines an inner product on the tangent space $T_p\mathcal{M}$. Concretely, g_p is symmetric and positive-definite:

$$g_p(v,v) \geq 0 \quad \text{for all } v \in T_p\mathcal{M} \quad \text{and} \quad g_p(v,v) = 0 \Leftrightarrow v = 0.$$

By identifying g_p with an inner product, we write

$$\langle v, w \rangle_g := g_p(v,w) \quad \text{for } v, w \in T_p\mathcal{M}.$$

A *Riemannian manifold* is then the pair $(\mathcal{M}, g)$.

Let $(x^1, \ldots, x^n)$ be local coordinates near $p \in \mathcal{M}$. Then, the basis for $T_p\mathcal{M}$ is $\left(\frac{\partial}{\partial x^1}\big|_p, \ldots, \frac{\partial}{\partial x^n}\big|_p\right)$. Tangent vectors $v, w \in T_p\mathcal{M}$ are expressed as $v = \sum_{i=1}^n v^i \frac{\partial}{\partial x^i}\big|_p$ and $w = \sum_{i=1}^n w^i \frac{\partial}{\partial x^i}\big|_p$, respectively. The matrix notation G_p of g at p consists of (i,j)-elements

$$g_{ij}(p) = g_p\left(\frac{\partial}{\partial x^i}\bigg|_p, \frac{\partial}{\partial x^j}\bigg|_p\right) = \left\langle \frac{\partial}{\partial x^i}\bigg|_p, \frac{\partial}{\partial x^j}\bigg|_p \right\rangle_g \tag{10}$$

for $i, j = 1, 2, ..., n$. The Euclidean metric is written as an identity matrix I. The inner product of v and w with respect to the Riemannian metric g_p can be expressed as:

$$g_p(v,w) = \sum_{i,j=1}^n g_{ij}(p) v^i w^j = v^T G_p w. \tag{11}$$

Lengths of Tangent Vectors and Curves. Given $\langle v, v \rangle_g$, the length of a tangent vector $v \in T_p\mathcal{M}$ is given by $|v|_g := \sqrt{\langle v, v \rangle_g}$. For a smooth curve $\gamma : [0,1] \to \mathcal{M}$, its length is defined by

$$L(\gamma) := \int_0^1 |\gamma'(t)|_g dt = \int_0^1 \sqrt{\langle \gamma'(t), \gamma'(t) \rangle_g} dt = \int_0^1 \sqrt{\gamma'(t)^\top G_{\gamma(t)} \gamma'(t)} dt. \tag{12}$$

For convenience, we denote this integrand of Eq. (12) as:

$$l(t) := \sqrt{\gamma'(t)^\top G_{\gamma(t)} \gamma'(t)}. \tag{13}$$

B COMPARISON METHODS

B.1 DDIM INVERSION

Naive encoding of an original image is simply adding Gaussian noise as in the forward process $q(x_t|x_{t-1})$, which is stochastic and often yields poor reconstructions. To accurately invert the reverse process and recover the specific noise map associated with a given image, *DDIM Inversion* (Mokady et al., 2023) is widely used. The key insight is that, in the limit of infinitesimally small time steps, the ODE formulation of DDIM is invertible.

Concretely, setting $\sigma_t = 0$ in Eq. (7) gives

$$x_{t-1} = a_t x_t + b_t \epsilon_\theta(x_t, t), \tag{14}$$

where $a_t = \sqrt{\alpha_{t-1}/\alpha_t}$ and $b_t = -\sqrt{\alpha_{t-1}(1-\alpha_t)/\alpha_t} + \sqrt{1-\alpha_{t-1}}$. With a sufficiently small time step size,

$$x_t = \frac{x_{t-1} - b_t \epsilon_\theta(x_t, t)}{a_t} \approx \frac{x_{t-1} - b_t \epsilon_\theta(x_{t-1}, t)}{a_t}, \tag{15}$$

as $\epsilon_\theta(x_t, t) \approx \epsilon_\theta(x_{t-1}, t)$. Iteratively applying the update rule in Eq. (15) to a sample x_0 from $t = 1$ to τ recovers the noisy image x_τ that would generate the original x_0. This inversion procedure substantially improves the fidelity of reconstructions and subsequent interpolations.

B.2 LINEAR INTERPOLATION

Once the noisy images are recovered via DDIM Inversion, one can perform straightforward linear interpolation (Lerp) (Song et al., 2021a), by treating the noise space (the data space with $t > 0$) as a linear latent space. In particular, let $x_\tau^{(0)}$ and $x_\tau^{(1)}$ denote the noisy versions of $x_0^{(0)}$ and $x_0^{(1)}$ in the noise space at $t = \tau$, respectively. A linear interpolation in that space is given by

$$x_\tau^{(s)} = (1-s)x_\tau^{(0)} + s x_\tau^{(1)}, \tag{16}$$

where $s \in [0, 1]$ is the interpolation parameter. Then, one then applies the reverse process from $t = \tau$ back to $t = 0$ to obtain the interpolated images $x_0^{(s)}$ in the data space.

B.3 SPHERICAL LINEAR INTERPOLATION

An alternative is spherical linear interpolation (Slerp) (Shoemake, 1985), which finds the shortest path on the unit sphere in the noise space:

$$x_\tau^{(s)} = \frac{\sin((1-s)\theta)}{\sin(\theta)} x_\tau^{(0)} + \frac{\sin(s\theta)}{\sin(\theta)} x_\tau^{(1)} \tag{17}$$

where $\theta = \arccos\left(\frac{(x_\tau^{(0)})^\top x_\tau^{(1)}}{\|x_\tau^{(0)}\|\|x_\tau^{(0)}\|}\right)$. Because this procedure preserves the norms of the noisy images $x_\tau^{(s)}$, it often yields natural interpolations than Lerp. Note that, Slerp assumes that $x_\tau^{(0)}$ and $x_\tau^{(1)}$ are drawn from a normal distribution, which holds only for a sufficiently large τ. Nonetheless, Slerp typically performs better with moderate τ.

B.4 PROPOSED METHOD

The velocities $v_t^{(i)}$ are approximated using second-order finite differences:

$$v_t^{(i)} = \begin{cases} \frac{-3x_t^{(0)} + 4x_t^{(1)} - x_t^{(2)}}{2\Delta s} & (i = 0) \\ \frac{x_t^{(i+1)} - x_t^{(i-1)}}{2\Delta s} & (0 < i < N) \\ \frac{3x_t^{(N)} - 4x_t^{(N-1)} + x_t^{(N-2)}}{2\Delta s} & (i = N) \end{cases} \tag{18}$$

C ADDITIONAL EXPERIMENTS AND RESULTS

C.1 CLIP-IQA

CLIP-IQA (Wang et al., 2023) is a metric that evaluates the quality of images generated by generative models. It is based on a pre-trained language-image model, CLIP (Radford et al., 2021), which predicts the similarity between images and text. For evaluating the reality of images, CLIP-IQA uses a pair of prompts: "Real photo" and "Abstract photo," and evaluate how similar the generated images are to the prompts. If it is close to the "Real photo" prompt, the score approaches 1.0, and the image is considered realistic. For noisiness, the prompts are "Clean photo" and "Noisy photo".

While this goes beyond CLIP-IQA's original scope, we used the prompts "A photo of [object]" and "A photo of something that is not [object]" to evaluate the fidelity of interpolated images to the prompt.

C.2 EXPERIMENTS WITH MNIST

We also evaluated our method with MNIST (Deng, 2012), as well as Lerp (Song et al., 2021a), Slerp (Shoemake, 1985). We set the maximum number of forward steps to $T = 1000$ and the number of forward steps before interpolation to $\tau = 400$. We set the number of interpolation steps to $N = 10$. We optimized the geodesic path using Adam (Kingma & Ba, 2015) for 5,000 iterations. The learning rate was initialized at 10^{-3}.

Unlike the results from Stable Diffusion (Figure 1), Lerp produces interpolated results with less noisy than Slerp. This suggests that the MNIST data manifold is relatively locally linear. However, Lerp exhibits discontinuous changes in digits; for example, a sudden appearance of a "9" during

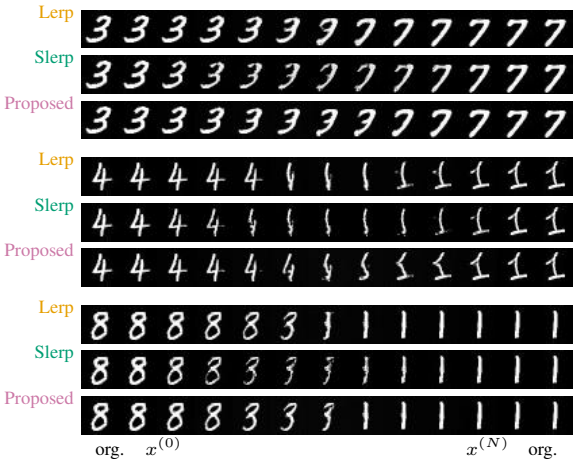

Figure 2: Interpolation results by a diffusion model trained on MNIST.

the transition from “3” to “7”. This is likely because Lerp ignores the underlying metric of the data manifold. In contrast, the proposed method shows gradual transitions compared to both Lerp and Slerp, achieving geometrically consistent transitions.