Be Tangential to Manifold: Discovering Riemannian Metric for Diffusion Models

Shinnosuke Saito[1] and Takashi Matsubara[2,1]

[1]Hokkaido University, Japan [2]AI Lab, CyberAgent, Inc., Japan
saitou.shinnosuke.y0@elms.hokudai.ac.jp
matsubara_takashi@cyberagent.co.jp

Abstract. Diffusion models are powerful deep generative models, but unlike classical models, they lack an explicit low-dimensional latent space that parameterizes the data manifold. This absence makes it difficult to perform manifold-aware operations, such as geometrically faithful interpolation or conditional guidance that respects the learned manifold. We propose a training-free Riemannian metric on the noise space, derived from the Jacobian of the score function. The key insight is that the spectral structure of this Jacobian separates tangent and normal directions of the data manifold; our metric leverages this separation to encourage paths to stay tangential to the manifold rather than drift toward high-density regions. To validate that our metric faithfully captures the manifold geometry, we examine it from two complementary angles. First, geodesics under our metric yield perceptually more natural interpolations than existing methods on synthetic, image, and video frame datasets. Second, the tangent–normal decomposition induced by our metric prevents classifier-free guidance from deviating off the manifold, improving generation quality while preserving text-image alignment.

Keywords: Diffusion Models · Manifold Hypothesis · Riemannian Geometry · Image Interpolation · Classifier-Free Guidance

1 Introduction

Diffusion models are a class of deep generative models (DGMs) that have shown a remarkable capability to generate high-fidelity, diverse content [33, 73, 81]. A key theoretical lens for understanding and improving DGMs is the *manifold hypothesis*, which states that real-world data (e.g., images) are concentrated around a low-dimensional manifold embedded in the high-dimensional data space [8, 20]. Under this hypothesis, DGMs are understood to learn not only the data distribution but also its underlying manifold, either explicitly or implicitly [54].

Among DGMs, variational autoencoders (VAEs) [46] and generative adversarial networks (GANs) [24] possess a low-dimensional latent space, which serves as an explicit parameterization of the data manifold [3]. On this latent space, one can define a Riemannian metric by pulling back the data-space metric through the decoder [25], enabling geometrically meaningful operations. For example,

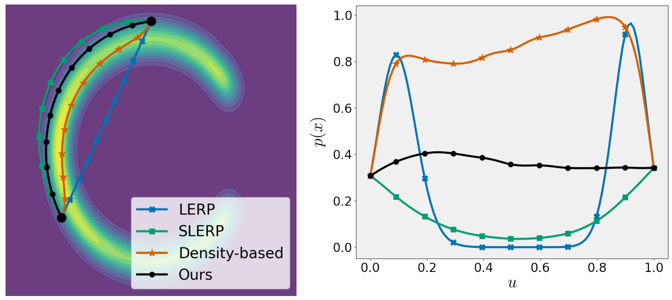

Fig. 1: A prominent example. (*left*) Interpolation paths on a C-shaped distribution. (*right*) A plot of the probability density transitions for interpolation paths. LERP cuts through a low-density region. SLERP deviates from the manifold. Density-based interpolation approaches a high-density region and traverses it, not preserving the probabilities of the endpoints. **Ours** runs parallel to the manifold, yielding natural transitions with preserved endpoint probabilities.

traversing the latent space along geodesics yields interpolations that are faithful to the intrinsic geometric structure of the data [5, 6, 13, 75]. However, diffusion models lack such a low-dimensional latent space; their noise space has the same dimensionality as the data space, making it nontrivial to define a meaningful pullback metric.

Some previous studies have attempted to define a Riemannian metric for diffusion models based on the density of noisy samples or the direction to the manifold (see Fig. 1) [7, 98]. Geodesics under these metrics, however, may overly approach the center of the data distribution, resulting in over-smoothed images [39]. The fundamental issue is that density or direction only tell us *where* samples concentrate, but not *how* the manifold is oriented in the ambient space. This raises a natural question: *can we define a Riemannian metric for diffusion models that directly reflects the intrinsic geometry of the data manifold, rather than its density?*

We answer this affirmatively. We build on prior observations that the Jacobian of the score function already encodes this geometry: its spectral structure separates the tangent and normal directions of the data manifold (illustrated conceptually in Fig. 2) [85, 91]. Building on this insight, we propose a Riemannian metric on the noise space, defined as the pullback of the Euclidean metric through the score function. This metric assigns high cost to movement along normal directions and low cost along tangent directions, thereby encouraging paths to stay tangential to the data

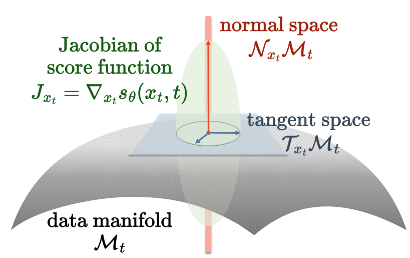

Fig. 2: A conceptual illustration. Our metric is derived from the Jacobian J_{x_t} of the score function s_θ, whose spectral structure separates tangent and normal directions, $\mathcal{T}_{x_t}\mathcal{M}_t$ and $\mathcal{N}_{x_t}\mathcal{M}_t$, of the data manifold $\mathcal{M}_t$.

manifold. Because it is constructed solely from the Jacobian of a pre-trained score function, it requires no additional training or architectural modifications.

The contribution of this work is the proposed Riemannian metric itself, which turns this known tangent–normal characterization into a concrete geometric foundation for understanding the data manifold in diffusion models and designing manifold-aware operations. We examine the metric from two complementary angles. Together, these angles serve to validate that the metric faithfully captures the geometry of the data manifold and to demonstrate its practical utility. **(i) Global geometry through interpolation** (Secs. 4.1 to 4.3): Geodesics under our metric should stay on or run parallel to the manifold, producing perceptually natural transitions. Experiments on synthetic, image, and video frame datasets confirm this. **(ii) Local geometry through guidance correction** (Sec. 4.4): The tangent–normal decomposition induced by our metric should prevent classifier-free guidance (CFG) [34] from pushing samples off the manifold. The guidance correction consistently improves FID while maintaining CLIP Score, and the improved quality transfers to distilled models at no additional overhead.

2 Related Work

Riemannian Geometry of Deep Generative Models. In VAEs and GANs, the latent space parameterizes the data manifold, and linear traversals in this space are widely used for editing semantic attributes of generated images [23, 27, 30, 61, 67, 77, 78, 84, 92, 105]. However, as real-world data distributions are skewed and heterogeneous, linear manipulations often fall short, motivating nonlinear approaches [1, 2, 14, 15, 37, 43, 52, 71, 89, 90]. A principled way to handle such nonlinearity is to equip the latent space with a Riemannian metric [25]. Some methods learn a metric through additional networks, which may overwrite the original manifold structure [4,50,83,95]. Others construct the pullback metric from the Euclidean metric on the data space through the decoder (or the generator) of a pre-trained model without additional training [5, 6, 13, 18, 75]. Our work follows the latter, training-free approach, but extends it to diffusion models, whose noise space is not low-dimensional.

Data Manifold in Diffusion Models. Diffusion models learn a score function $s_\theta(x_t, t)$, which iteratively denoises noisy samples backward in time from $t = T$ to $t = 0$ to form the data distribution [33,73,80–82]. A space of noisy samples at $t > 0$ is often referred to as a *noise space*. Diffusion models have been shown to implicitly learn the data manifold through this denoising process [22, 65, 68, 88, 94, 99]. One common assumption is that high-density regions of the learned distribution correspond to the data manifold, but recent studies have shown that image likelihood is negatively correlated with perceptual detail: images in high-density regions are often over-smoothed, whereas those in lower-density regions may contain richer textures [39]. From a different perspective, several studies have estimated the local intrinsic dimension of the data manifold [35, 36, 38, 85, 91]. Their key insight is that the Jacobian of the score function exhibits a rank deficiency corresponding to the codimension of the data manifold, or in practice a sharp spectral gap, which separates the normal directions of the data manifold from the tangent directions [85, 91]. Rather than relying on density, we build upon this spectral characterization to define a Riemannian metric on the noise space of a pre-trained diffusion model.

Interpolation in Diffusion Models. Although latent diffusion models [73] operate on the latent space of an autoencoder, this space is designed for perceptual compression rather than semantic representation, and simple interpolations therein do not yield coherent transitions. Interpolation methods in the noise space can be broadly categorized into four groups. (i) Closed-form methods operate directly in the noise space. LERP [33] treats it as a flat Euclidean space, leading to small norms and loss of detailed features. SLERP [79, 81] preserves norms along a hypersphere, and other variants adjust norms by other heuristics [10, 74, 102]. (ii) Surrogate latent-space methods exploit intermediate layers of the diffusion network [31, 47, 62, 63], but skip connections allow information to bypass these layers, limiting their ability to serve as self-contained representations. (iii) Training-dependent methods employ additional networks or modify the architecture for interpolation [26, 28, 44, 58, 69, 76, 93, 96, 100], but require additional training and cannot be applied to arbitrary pre-trained models. (iv) Riemannian metric-based methods define a geometry on the noise space without additional training. GeodesicDiffusion [98] defines a conformal metric using the inverse density of noisy samples, and Azeglio and Bernardo [7] propose a metric inspired by the Fisher information matrix (FIM); both guide geodesics toward high-density regions, as examined for normalizing flows [83] and energy-based models [9]. However, high-density regions do not necessarily correspond to perceptually detailed images [39], which can result in over-smoothed and cartoonish interpolations. Although some studies draw inspiration from statistical manifolds, it remains unclear what geometric structures their methods leverage [40, 55]. Our metric falls in this category but differs fundamentally: rather than favoring high-density regions, it leverages the spectral structure of the score Jacobian to keep geodesics tangential to the data manifold, keeping the density level.

Guidance Methods. One can condition the denoising process on a text prompt [73]. Classifier-free guidance (CFG) amplifies this condition to make generated images more faithful to text prompts [34], and the negative prompt suppresses concepts specified by a complementary prompt [73]. Some guidance methods additionally utilize external information [12, 72, 86]. However, it is well known that these guidance methods can distort the learned data manifold, leading to visually unnatural outcomes. Recent studies have attempted to mitigate this issue. CFG++ [17] uses the unconditional score instead of the CFG-based score for renoising steps, while TCFG [48] projects the unconditional score onto the subspace shared with the conditional one via SVD. Although these methods are motivated by the existence of the data manifold, they do not formally define or characterize its geometry, and their corrections remain heuristic. Our metric provides such an explicit geometric characterization, which enables a geometrically grounded correction that decomposes the guidance term into tangent and normal components and suppresses the latter.

3 Method

3.1 Proposed Riemannian Metric

Definition of Proposed Metric. We define a Riemannian metric on the noise space that captures the local geometry of the data manifold through the score function. See Appendix A.1 for background. Let x_t be a point in the noise space $\mathbb{R}^D$ at time t, and $v, w \in T_{x_t}\mathbb{R}^D$ be tangent vectors to $\mathbb{R}^D$ at x_t. Let $s_\theta(x_t, t)$ be the score function of a pre-trained diffusion model parameterized by θ. We propose the following Riemannian metric at time t:

$$g_{x_t}(v, w) \coloneqq \langle J_{x_t} v, J_{x_t} w \rangle = v^\top J_{x_t}^\top J_{x_t} w = v^\top G_{x_t} w, \tag{1}$$

where $J_{x_t} = \nabla_{x_t} s_\theta(x_t, t)$ is the Jacobian of the score function $s_\theta(\cdot, t)$ at x_t, and $G_{x_t} = J_{x_t}^\top J_{x_t}$ is the matrix representation of the metric g_{x_t}. This metric is the pullback $s_\theta^* I$ of the Euclidean metric I on the score space $\mathbb{R}^D$ through the score function s_θ, since $v^\top G_{x_t} w = (J_{x_t} v)^\top I (J_{x_t} w)$. By construction, G_{x_t} is symmetric and positive semi-definite; we show below that it is positive definite in practice.

Interpretation. Stanczuk et al. [85] found that the score function $s_\theta(x_t, t)$ points orthogonally towards the data manifold $\mathcal{M}_t$ containing the data point x_t, and that this orthogonality becomes more prominent as t approaches zero. Ventura et al. [91] further investigated the Jacobian J_{x_t} and observed that its rank deficiency (for clean samples on a low-dimensional manifold) or spectral gap (for real-world data) reflects the intrinsic dimension of the data manifold. Intuitively, J_{x_t} shrinks along tangent directions and remains large along normal directions, as shown in Fig. 2.

More precisely, let $\mathcal{M}_t$ be the data manifold at time t learned by a diffusion model, and $x_t \in \mathcal{M}_t$ be a point on the data manifold. Define the tangent space $\mathcal{T}_{x_t}\mathcal{M}_t$ as the d-dimensional subspace ($d \ll D$) spanned by the right singular vectors of J_{x_t} corresponding to small singular values; the normal space $\mathcal{N}_{x_t}\mathcal{M}_t$ is the orthogonal complement spanned by those corresponding to large singular values. Then, $T_{x_t}\mathbb{R}^D = \mathcal{T}_{x_t}\mathcal{M}_t \oplus \mathcal{N}_{x_t}\mathcal{M}_t$. From Eq. (1), the squared metric norm of a tangent vector v to the ambient noise space $\mathbb{R}^D$ is $g_{x_t}(v, v) = \|J_{x_t} v\|_2^2$.

Proposition 1. *Among tangent vectors v of a fixed Euclidean norm, those with the smallest squared metric norm $g_{x_t}(v, v) = \|J_{x_t} v\|_2^2$ lie in the tangent space $\mathcal{T}_{x_t}\mathcal{M}_t$ to the data manifold $\mathcal{M}_t$.*

Consequently, our metric assigns shorter lengths to paths that stay on or run parallel to the data manifold $\mathcal{M}_t$. See Appendix B.1 for a detailed derivation.

Geodesics. A geodesic is a shortest path under a given metric, obtained by minimizing the energy functional $E[\gamma]$ (see Appendix A.1 and Lee [49] for details). Let $u \in [0,1]$ parameterize a curve $\gamma : u \mapsto \gamma(u)$. The energy functional with our metric is:

$$\begin{aligned} E[\gamma] &= \tfrac{1}{2}\int_0^1 g_{\gamma(u)}(\gamma'(u), \gamma'(u))\mathrm{d}u = \tfrac{1}{2}\int_0^1 \langle J_{\gamma(u)}\gamma'(u), J_{\gamma(u)}\gamma'(u)\rangle \mathrm{d}u \\ &= \tfrac{1}{2}\int_0^1 \|J_{\gamma(u)}\gamma'(u)\|_2^2 \mathrm{d}u = \tfrac{1}{2}\int_0^1 \|\tfrac{\partial}{\partial u} s_\theta(\gamma(u), t)\|_2^2 \mathrm{d}u. \end{aligned} \tag{2}$$

This expression admits two complementary readings. The second-last term shows that geodesics minimize the squared metric norm $\|J_{x_t} v\|_2^2$ of the velocity $v = \gamma'(u)$ at each point $x_t = \gamma(u)$ along the path, which by Proposition 1 encourages the path to stay tangential to the data manifold. The last term, obtained by the chain rule, shows that geodesics equivalently minimize the squared change in the score function s_θ along the path. Since our metric is the pullback metric through s_θ, geodesics are mapped to straight lines in the score space. Because the score function encodes the gradient of the log-density, a path with minimal score variation maintains a consistent relationship with the density landscape: it neither approaches nor deviates from high-density regions, but runs parallel to the manifold at a consistent distance. This provides a geometric explanation for why our geodesics preserve endpoint probabilities (Fig. 1 (right)). Moreover, considering that gradients of log-likelihoods with respect to model parameters have been used as semantic representations of inputs [11, 29, 97], geodesics that minimize score variation can also be viewed as preserving semantic closeness between samples.

Given a geodesic, the distance between its endpoints is the length of the geodesic, computed as $d_g(\gamma(0), \gamma(1)) = \int_0^1 \sqrt{g_{\gamma(u)}(\gamma'(u), \gamma'(u))}\mathrm{d}u$.

Theoretical Remarks. Our metric $G_{x_t} = J_{x_t}^\top J_{x_t}$ captures the full spectral structure of the Jacobian, that is, all tangent and normal directions. This contrasts with the FIM-based metric $\lambda s_\theta s_\theta^\top + I$ [7], which captures only the rank-1 direction of the score s_θ and thus respects only one normal direction. The distinction is essential when the codimension of the data manifold is greater than one, which is the typical case for real-world data (e.g., images encoded in a representation space of tens of thousands of dimensions exhibit intrinsic dimensionality of a few hundred [85, 91]). Thus, practical diffusion models operate in a regime where the tangent–normal distinction is meaningful.

Diffusion models learn the score function s_θ directly rather than the log-density $\log p_t$, and consequently, J_{x_t} is not exactly symmetric. Even then, J_{x_t} typically exhibits a spectral gap, and the decomposition into $\mathcal{T}_{x_t}\mathcal{M}_t \oplus \mathcal{N}_{x_t}\mathcal{M}_t$ in Proposition 1 holds approximately.

The Jacobian J_{x_t} is degenerate on ideally clean data lying on a low-dimensional manifold at $t = 0$, but for real-world data it is full-rank with a sharp spectral gap. In practice, operations are performed at an intermediate time $t > 0$, where samples are corrupted by noise and the Jacobian J_{x_t} is full-rank and exhibits a moderate spectral gap, ensuring that g_{x_t} is positive definite and thus a proper Riemannian metric.

Algorithm 1 Interpolation

Require: Samples $x_0^{(0)}$, $x_0^{(1)}$, score s_θ, time τ, points N, iterations K
$\{x_\tau^{(0)}, x_\tau^{(1)}\} \leftarrow \{\text{DDIM-Fwd}(x_0^{(i)}, \tau)\}_{i\in\{0,1\}}$
$\{x_\tau^{(u_i)}\}_{i=1}^{N-1} \leftarrow \{\text{SLERP}(x_\tau^{(0)}, x_\tau^{(1)}, u_i)\}_{i=1}^{N-1}$
for $k = 1$ **to** K **do**
  Compute E via Eq. (3)
  Update $\{x_\tau^{(u_i)}\}_{i=1}^{N-1}$ to minimize E
end for
$\{\hat{x}_0^{(u_i)}\}_{i=0}^{N} \leftarrow \{\text{DDIM-Bwd}(x_\tau^{(u_i)}, \tau)\}_{i=0}^{N}$
return $\{\hat{x}_0^{(u_i)}\}_{i=0}^{N}$

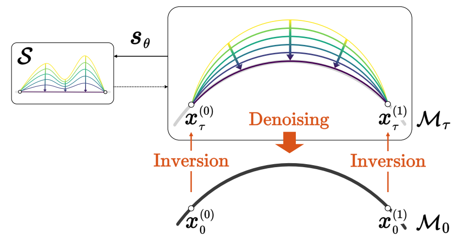

Fig. 3: Illustration of geodesic-based interpolation

See Appendix B.1 for further details, including a discussion of regularization and alternative construction.

3.2 From Metric to Algorithms

Geodesic-Based Interpolation. To use our metric for interpolation, we compute a geodesic in the noise space at a fixed time $t = \tau > 0$ by discretizing the path and minimizing the resulting energy. We approximate the path γ as $N+1$ points $x_\tau^{(u_i)}$ with $u_0 = 0$, $u_N = 1$, $\Delta u = 1/N$, and $x_\tau^{(u_i)} = \gamma(u_i)$. The energy functional $E[\gamma]$ in Eq. (2) is then approximated as:

$$E[\gamma] \approx \frac{1}{2\Delta u} \sum_{i=0}^{N-1} \|(s_\theta(x_\tau^{(u_{i+1})}, \tau) - s_\theta(x_\tau^{(u_i)}, \tau))\|_2^2. \tag{3}$$

Given two samples $x_\tau^{(0)}$ and $x_\tau^{(1)}$ at time τ, the geodesic is obtained by minimizing Eq. (3) with respect to the intermediate points $x_\tau^{(u_1)}, \dots, x_\tau^{(u_{N-1})}$. The distance is then $d_g(x_\tau^{(0)}, x_\tau^{(1)}) \approx \sum_{i=0}^{N-1} \|(s_\theta(x_\tau^{(u_{i+1})}, \tau) - s_\theta(x_\tau^{(u_i)}, \tau))\|_2$.

This follows common practice in diffusion models, where interpolation is performed in the noise space rather than directly in the data space at $t = 0$ (see Algorithm 1 and Fig. 3). Given a pair of clean samples $x_0^{(0)}$ and $x_0^{(1)}$, we first map them to noisy samples $x_\tau^{(0)}$ and $x_\tau^{(1)}$ at the fixed time τ using DDIM Inversion [60]. We then compute the geodesic path between them by minimizing Eq. (3), and map the interpolated noisy samples $x_\tau^{(u)}$ for $u = u_0, \dots, u_N$ back to clean samples $x_0^{(u)}$ using the deterministic denoising process (see Appendix A.2). Here, $x_0^{(u_1)}, \dots, x_0^{(u_{N-1})}$ serve as interpolated samples. The optimization is more computationally expensive than closed-form methods such as LERP or SLERP, but comparable to other metric-based methods [7, 98] that also require solving an optimization problem, as summarized in Appendices C.1 and C.2.

Metric-Based Guidance Correction. Complementary to global geodesics, our metric also provides a local tangent–normal decomposition at each point in the noise space. This decomposition provides a metric-based way to correct perturbations that push a sample off the data manifold. We apply this to classifier-free guidance (CFG), where the guidance term is known to distort the learned manifold [17, 48] (see Algorithm 2 and Fig. 4).

Algorithm 2 Guidance Correction

```
Require: x_T ~ N(0, I), condition c, CFG scale
    w, weight λ
for t = T to 1 do
    s ← s_θ(x_t, t, c)
    s̃ ← s̃_θ(x_t, t, c, w)                ▷ Eq. (17)
    Δs ← s̃ − s
    Compute Δŝ* via Eq. (5)
    ŝ ← s + Δŝ*
    x_{t−1} ← DDIMStep(x_t, ŝ)
end for
return x_0
```

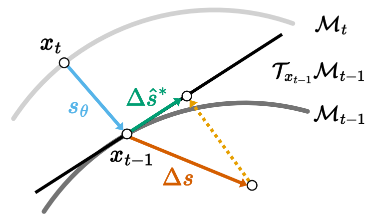

Fig. 4: Illustration of metric-based guidance correction

Diffusion models generate samples by iteratively updating a noisy sample x_t as $x_{t-1} = x_t + \eta_t s_\theta(x_t, t)$, where the noise term is omitted for simplicity, and η_t is the update size depending on time t. CFG modifies this by adding a guidance term $\Delta s(x_t, c)$ to $s_\theta(x_t, t)$, where c is some condition. Our correction adjusts this guidance term to balance two objectives: the corrected sample should remain close to the guided sample in Euclidean distance, while staying close to the unguided sample under our metric g, i.e., on the data manifold. Formally, we solve:

$$\Delta\hat{s}^* = \arg\min_{\Delta\hat{s}} d_E(x_t + \eta_t(s_\theta + \Delta s),\ x_t + \eta_t(s_\theta + \Delta\hat{s}))^2 + \lambda d_g(x_t + \eta_t s_\theta,\ x_t + \eta_t(s_\theta + \Delta\hat{s}))^2, \tag{4}$$

where d_E is the Euclidean distance.

For tractability, we evaluate d_g at time t rather than $t-1$ (valid because the manifold changes smoothly between adjacent timesteps) and approximate the path integral with $N = 1$ (sufficient for a local correction). Absorbing η_t into λ, the corrected guidance $\Delta\hat{s}^*$ is then given by $(I + \lambda G_{x_t})\Delta\hat{s}^* = \Delta s$. We solve this by the conjugate gradient (CG) method with Δs as the initial point, using only a single update step to limit computational overhead:

$$\Delta\hat{s}^* = \Delta s + \tfrac{\epsilon^\top \epsilon}{\epsilon^\top (I + \lambda G_{x_t})\epsilon}\epsilon \text{ for } \epsilon = \Delta s - (I + \lambda G_{x_t})\Delta s. \tag{5}$$

Since $G_{x_t} = J_{x_t}^\top J_{x_t}$, multiplication by G_{x_t} can be decomposed into a Jacobian-vector product followed by a vector-Jacobian product. We approximate the Jacobian-vector product $J_{x_t} v$ via a finite difference $(s_\theta(x_t + hv, t) - s_\theta(x_t, t))/h$ for a small $h > 0$, and approximate the vector-Jacobian product $J_{x_t}^\top v$ in the same way under the empirically supported assumption that J_{x_t} is approximately symmetric. In total, the correction requires four additional evaluations of s_θ. With CFG or negative prompt, each denoising step involves two evaluations of s_θ, so the computational cost is approximately three times that without correction. Since our correction modifies only the guidance term, it can be applied during distillation, so that the improved quality is inherited by the student model at no additional inference cost.

4 Experiments

Our primary contribution is the Riemannian metric defined in Eq. (1). To validate that it faithfully captures the geometry of the data manifold, we examine its global properties through interpolation (Secs. 4.1 to 4.3) and its local properties through guidance correction (Sec. 4.4).

4.1 Synthetic 2D Data

To illustrate the behavior of geodesics under our metric, we conducted experiments on a synthetic C-shaped distribution in 2D (see Fig. 1 and Appendix D.1 for details). We trained a DDPM [33] with a SiLU MLP backbone [19] on 100,000 samples from this distribution with $T = 50$ steps and obtained interpolations at $\tau = 0.02T = 1$ via DDIM Inversion. As shown in Fig. 1 (left), LERP traverses through low-density regions, SLERP slightly deviates from the manifold, and density-based interpolation [98] approaches a high-density region, not preserving the endpoint probabilities. In contrast, our geodesic runs parallel to the data manifold, preserving the endpoint probabilities.

To quantify this, we sampled 50 pairs of endpoints and evaluated the standard deviation of the density $p(x)$ along each interpolation path, summarized in Tab. 1. A smaller value indicates a more consistent distance from the data manifold. Our metric achieves the lowest standard deviation, which confirms that geodesics under our metric maintain a consistent relationship with the manifold, as predicted by the analysis in Sec. 3.

Table 1: Results on synthetic 2D data

Method	Std. ↓
LERP	0.1606
SLERP	<u>0.0833</u>
Density	0.1073
Ours	**0.0701**

4.2 Image Interpolation

Experimental Setup. Following prior work [74,98,102], we use Stable Diffusion v2.1-base [73] as the backbone with $T = 50$ timesteps and $N - 1 = 9$ interpolated images. We do not use CFG or negative prompts for DDIM inversion and denoising. For the geodesic computation, we use negative prompts and treat the resulting update $\tilde{s}_\theta$ (i.e., Eq. (18) in Appendix A.2) as the score function s_θ for computing the Jacobian J_{x_t} and thus the metric g_{x_t}. We evaluate on four benchmark datasets: the animation and metamorphosis subsets of MorphBench (MB(A) and MB(M)) [100], Animal Faces-HQ (AF) [16], and CelebA-HQ (CA) [41]. For AF and CA, we curate 50 image pairs with LPIPS [101] below 0.6 to ensure semantic similarity, following Yu et al. [98]. Further details are provided in Appendix D.2.

We compare against LERP [33], SLERP [81], NAO [74], NoiseDiffusion (NoiseDiff) [102], FIM-based metric [7], and GeodesicDiffusion (GeoDiff) [98]. We use default settings for NAO, NoiseDiff, and GeoDiff based on their official codes. Due to the lack of official code, we re-implemented the FIM-based metric with the same backbone as ours. For fair comparison, LERP, SLERP, FIM-based metric, and our proposed metric use the DDIM Scheduler [81] and operate in the noise space at $\tau = 0.6T$, following Yu et al. [98]. For FIM-based metric and our proposed metric, each path was initialized with SLERP and updated for 500 iterations using Adam optimizer [45] with a learning rate of 10^{-3}, decayed with cosine annealing to 10^{-4} [56]. We also adopted the prompt adjustment of Yu et al. [98], which optimizes the text embedding similarly to textual inversion [21]; see Appendices C.1 to C.3 for details on comparison methods, computational overhead, and prompt adjustment.

We evaluate interpolation quality using four measures. (1) Perceptual Path Length (PPL) [42] is the sum of LPIPS between adjacent images, assessing the directness of transitions. (2) Perceptual Distance Variance (PDV) [100] is the standard deviation of adjacent LPIPS, assessing the uniformity of transitions. (3) Fréchet Inception Distance (FID) [32] is a distributional distance between input and interpolated images via the features of Inception v3 [87]. (4) Reconstruction Error (RE) is the mean squared error between the input endpoints and the reconstructed endpoints. PPL, PDV, and FID are standard measures for image interpolation [74, 98, 102]; we additionally report RE to verify whether the endpoints are faithfully preserved. We follow the evaluation protocol of prior work, including the sample sizes for FID computation (900 interpolated images vs. 200 reference images on AF and CA). While none of these measures directly evaluates manifold faithfulness, consistent improvement across all four provides converging evidence. We further complement these indirect measures with ground-truth evaluation in video frame interpolation (Sec. 4.3).

Results. We summarize the quantitative results in Tab. 2. Our metric achieves the best scores on all datasets for PPL, FID, and RE, and records the best PDV on MB(A) with a close second-best on the others.

Qualitative results in Figs. 5 and 6 reveal further differences. LERP yields blurry interpolations. NAO and NoiseDiff produce vivid textures absent in the originals and exhibit large reconstruction errors, as their norm adjustments alter the endpoints. SLERP produces sharper results than LERP but lags behind geodesic-based methods. The most competitive baseline is GeoDiff, which ranks second on most metrics but produces over-smoothed images that lack fine details. This is consistent with the finding that sample density is negatively correlated with perceptual detail [39]: density-based metrics guide geodesics toward high-density regions, sacrificing texture. FIM-based metric yields cartoonish results with non-smooth transitions, as it is nearly singular (rank-1) with the default $\lambda = 1{,}000$. In contrast, our metric preserves fine details of the input images throughout the interpolation. We confirmed that these trends hold with a different backbone (Stable Diffusion v2.0-base); see Appendix E for complementary results, including ablation studies and visualizations of the spectral gap.

4.3 Video Frame Interpolation

Experimental Setup. To complement the indirect measures of the previous section, we evaluate methods on video frame interpolation using MSE and LPIPS

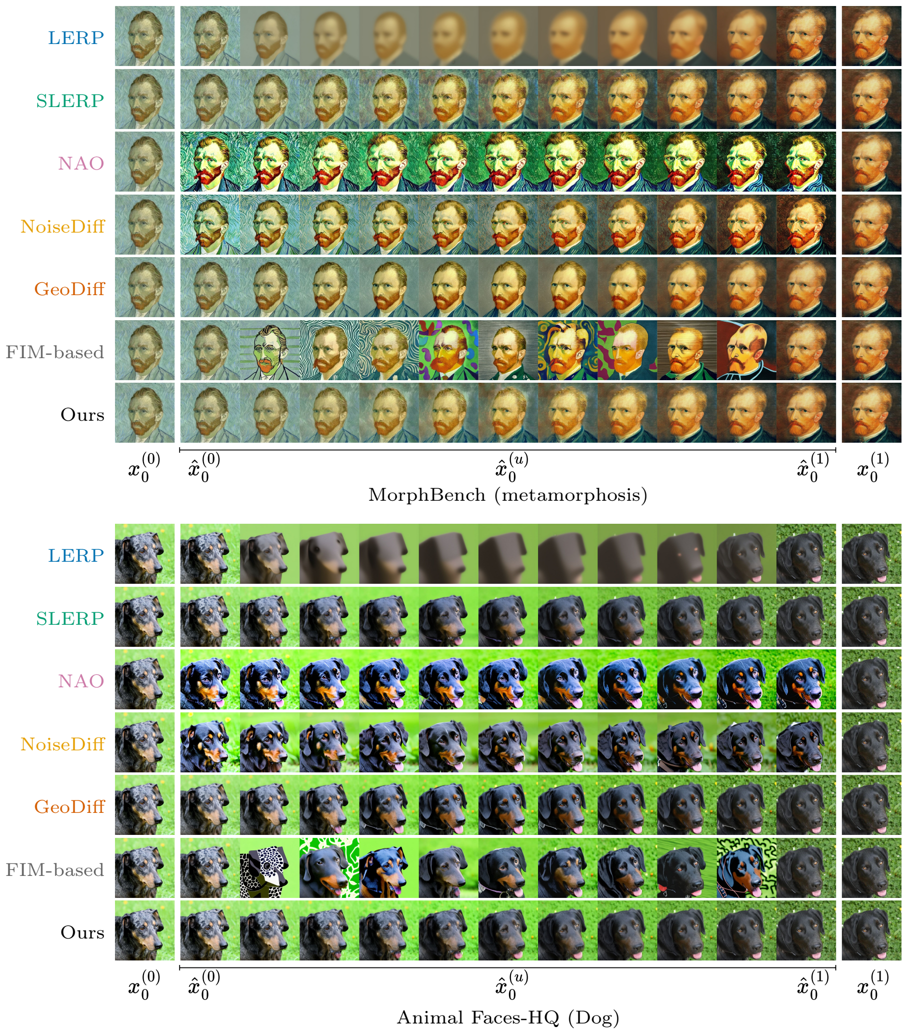

Fig. 5: Qualitative examples of interpolated image sequences for MB(M) and AF (Dog). The images at both ends are the given endpoints $x_0^{(0)}$ and $x_0^{(1)}$, and the middle images are the interpolated results $\{\hat{x}_0^{(u)}\}$ for $u \in [0, 1]$. See also Fig. E.1 in Appendix E.

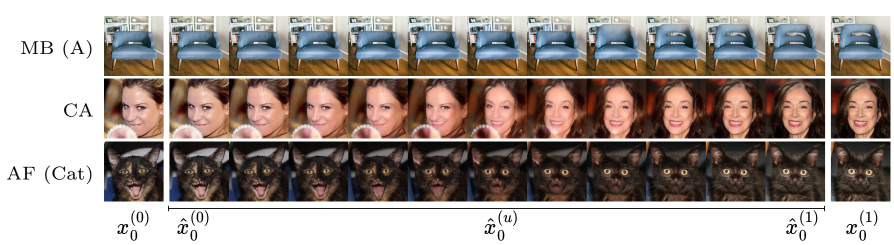

Fig. 6: Additional examples by our method. See also Fig. E.1 in Appendix E.

Table 2: Results on image interpolation

	PPL ↓				**PDV** ↓			
Method	MB(A)	MB(M)	CA	AF	MB(A)	MB(M)	CA	AF
LERP	0.848	1.787	1.420	1.859	0.055	0.128	0.091	0.154
SLERP	0.644	1.065	0.707	0.871	0.030	**0.055****	**0.033***	**0.022***
NAO	2.868	4.299	2.121	2.443	0.163	0.164	0.154	0.173
NoiseDiff	3.618	2.011	2.098	3.250	0.064	0.085	0.069	0.083
GeoDiff	0.402	1.021	0.669	0.842	0.024	0.073	0.044	0.027
FIM-based	3.358	4.429	4.152	5.249	0.142	0.187	0.172	0.196
Ours	**0.380***	**0.977***	**0.633****	**0.767****	**0.021***	0.073	0.036	0.023

	FID ↓				**RE** ↓ $(\times 10^{-3})$			
Method	MB(A)	MB(M)	CA	AF	MB(A)	MB(M)	CA	AF
LERP	84.20	118.90	95.68	119.58	0.401	0.397	1.010	2.049
SLERP	62.81	48.99	37.84	26.07	0.401	0.397	1.010	2.049
NAO	130.54	102.64	83.05	71.47	39.244	44.302	27.623	40.178
NoiseDiff	119.47	74.03	65.04	68.87	15.096	7.835	8.618	19.628
GeoDiff	28.70	38.12	35.98	25.80	0.188	0.272	0.891	1.969
FIM-based	92.09	78.80	70.95	59.11	0.401	0.397	1.010	2.049
Ours	**27.44**	**36.00**	**32.54**	**21.01**	**0.177***	**0.201****	**0.888***	**1.962****

* and ** indicate that the improvement over the second-best method is statistically significant at the 0.01 and 0.001 levels, respectively, according to a one-sided exact binomial test ($H_0:\ p = 0.5$).

against ground-truth middle frames. Our goal is not to compete with dedicated video interpolation methods, but to use ground-truth frames as an objective proxy for manifold faithfulness.

We employ three benchmarks curated by Zhu et al. [104]: 21 natural-scene clips from DAVIS [64], 56 human-pose clips from Pexels (Human), and 26 indoor/outdoor clips from RealEstate10K (RE10K) [103]. From each clip, we use three consecutive frames to ensure a unique interpolation target: frames 1 and 3 serve as $x_0^{(0)}$ and $x_0^{(1)}$, and frame 2 is used as the ground-truth middle frame for evaluating $\hat{x}_0^{(0.5)}$. Unless otherwise specified, all methods and hyperparameters are identical to those used for image interpolation. Each frame is resized to 512×512 pixels, and a text prompt is generated from frame 1 using BLIP-2 [51].

Results. Table 3 summarizes the quantitative results. Our method achieves the lowest MSE and LPIPS on all datasets, with statistically significant improvements over the second-best method. In Fig. 7, the qualitative trends observed in image interpolation carry over: LERP produces blurry outputs, NAO and NoiseDiff deviate substantially from the ground truth, and GeoDiff over-smooths textures. Notably, the zoomed-in comparison on Human shows that only our method and GeoDiff correctly interpolate the arm movement, but GeoDiff flattens water

Table 3: Results on video frame interpolation

	MSE ↓ (×10⁻³)			LPIPS ↓		
Method	DAVIS	Human	RE10K	DAVIS	Human	RE10K
LERP	12.135	4.566	6.299	0.590	0.379	0.377
SLERP	15.440	6.080	6.128	0.487	0.320	0.301
NAO	108.211	99.867	121.680	0.679	0.668	0.664
NoiseDiff	46.881	41.994	28.867	0.561	0.552	0.482
GeoDiff	13.253	3.363	5.941	0.334	0.184	0.229
FIM-based	30.172	11.638	12.679	0.535	0.388	0.373
Ours	**8.777****	**2.018****	**2.771****	**0.318***	**0.170****	**0.178****

* and ** indicate the statistical significance in the same manner as Tab. 2.

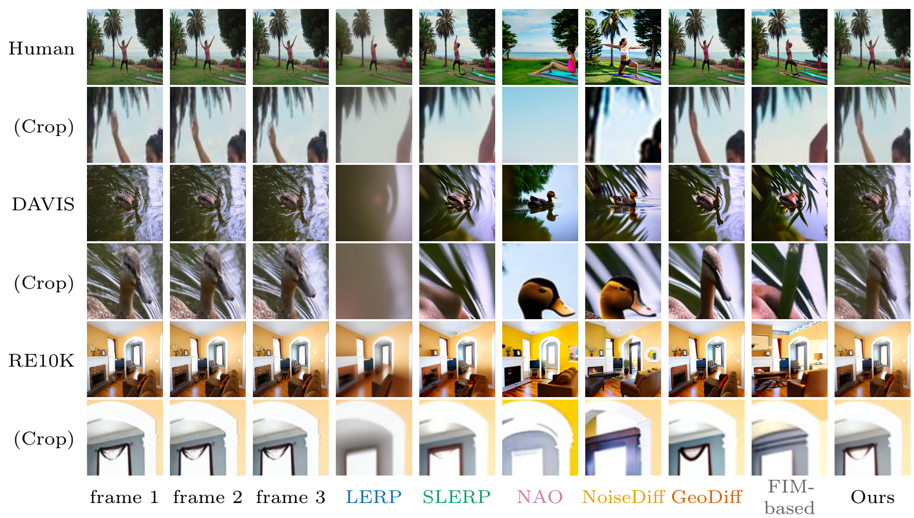

Fig. 7: Qualitative examples for video frame interpolation

ripples on DAVIS, confirming that density-based geodesics sacrifice fine detail. Our method preserves edges, object shapes, and textures most faithfully, as now confirmed by comparison against ground-truth frames.

4.4 Generation with Guidance Correction

Experimental Setup. As in the previous sections, we use Stable Diffusion v2.1-base [73] as the backbone with $T = 50$ timesteps. We generate 30k images using text prompts from the MS-COCO 2014 validation set [53] and evaluate using FID [32] against the corresponding images and CLIP Score [70] against the text prompts. The former evaluates image quality, while the latter evaluates text-image alignment. We use CFG [34] as the baseline and CFG++ [17] as a

Table 4: Results on guidance correction

	$w = 5.0$		$w = 7.5$		$w = 12.5$	
	FID ↓	CLIP ↑	FID ↓	CLIP ↑	FID ↓	CLIP ↑
CFG	11.69	0.313	14.29	0.314	17.28	0.315
CFG++	11.87	0.313	13.98	0.314	17.76	0.315
Ours	**11.53**	0.313	**13.81**	0.314	**16.04**	0.315

Table 5: Results on distillation

	FID ↓	CLIP ↑
CFG	17.91	0.306
Ours	**16.98**	0.306

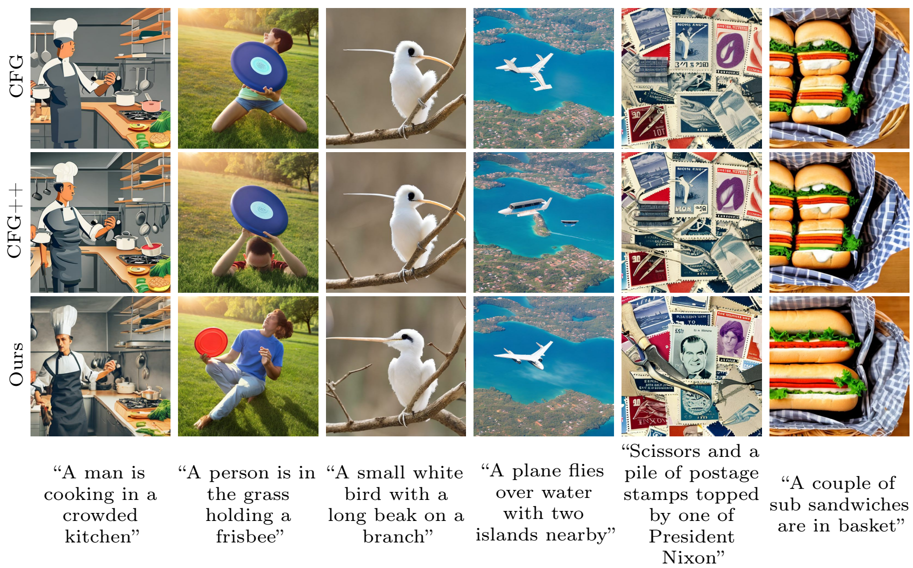

Fig. 8: Qualitative comparison between CFG, CFG++, and Ours

comparison method. We set the finite difference to $h = 10^{-4}$ and the weight to $\lambda = 0.1$ for our method and use the default hyperparameters of CFG++ from the original paper.

Results. We summarize the results in Tab. 4. Regardless of the CFG scale w, our metric-based guidance correction consistently improves FID while maintaining comparable CLIP Score. CFG++ does not consistently improve FID, suggesting sensitivity to hyperparameter settings. The fact that CLIP Score is preserved across all methods indicates that our guidance correction improves image quality without sacrificing text-image alignment, consistent with the design of our objective in Eq. (5).

Figure 8 shows representative examples using prompts from the same dataset. CFG and CFG++ produce cartoonish images in the leftmost column and overly emphasize some objects (i.e., frisbee and beak) in the second and third columns. Our guidance correction produces more natural images with better details, which confirms that it effectively prevents off-manifold perturbations back onto the data manifold.

Distillation with Guidance Correction. When our guidance correction modifies only the teacher's guidance term, it can be incorporated into distillation at no additional inference cost for the student. We follow the protocol of the latent consistency model (LCM) [59] with default settings in diffusers [66], including AdamW optimization [57]; see Appendix D.3 for details. After distillation, we generate images with $T = 4$ steps and evaluate them in the same way as above.

Table 5 summarizes the results averaged over five runs. For each run, we match the random seed between CFG and our method; our method improves FID in all five runs ($p < 0.05$, one-sided exact binomial test) while maintaining comparable CLIP Score.

5 Conclusion

We proposed a Riemannian metric on the noise space of diffusion models, derived from the Jacobian of the score function. Building on the spectral structure of this Jacobian, which separates tangent and normal directions of the data manifold, our metric encourages geodesics to stay within or run parallel to the manifold without any additional training or architectural modifications. We validated this metric from two complementary angles: geodesic-based interpolation confirmed that it captures global manifold geometry, and guidance correction confirmed that its local tangent–normal decomposition is accurate enough to project off-manifold perturbations back onto the manifold. Beyond these two tasks, we believe this metric provides a geometric foundation for broader manifold-aware operations in diffusion models, such as image editing, noise-space clustering, and extension to flow-matching and energy-based models. Validating the metric on more recent architectures remains important future work.

Acknowledgements

This study was partly supported by JST BOOST (JPMJBY24H0), JST CREST (JPMJCR24Q5), and JST ASPIRE (JPMJAP2329), Japan, and achieved through the use of large-scale computer systems at the D3Center, Osaka University.

References

1. Abdal, R., Zhu, P., Mitra, N.J., Wonka, P.: StyleFlow: Attribute-Conditioned Exploration of StyleGAN-Generated Images Using Conditional Continuous Normalizing Flows. ACM Transactions on Graphics (2021)
2. Aoshima, T., Matsubara, T.: Deep Curvilinear Editing: Commutative and Nonlinear Image Manipulation for Pretrained Deep Generative Model. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2023)

3. Arjovsky, M., Bottou, L.: Towards Principled Methods for Training Generative Adversarial Networks. In: International Conference on Learning Representations (ICLR) (2017)
4. Arvanitidis, G., Georgiev, B.M., Schölkopf, B.: A Prior-Based Approximate Latent Riemannian Metric. In: International Conference on Artificial Intelligence and Statistics (AISTATS) (2022)
5. Arvanitidis, G., Hansen, L.K., Hauberg, S.: Latent Space Oddity: on the Curvature of Deep Generative Models. In: International Conference on Learning Representations (ICLR) (2018)
6. Arvanitidis, G., Hauberg, S., Schölkopf, B.: Geometrically Enriched Latent Spaces. In: International Conference on Artificial Intelligence and Statistics (AISTATS) (2021)
7. Azeglio, S., Bernardo, A.D.: What's Inside Your Diffusion Model? A Score-Based Riemannian Metric to Explore the Data Manifold. arXiv (2025)
8. Bengio, Y., Courville, A.C., Vincent, P.: Representation Learning: A Review and New Perspectives. IEEE Transactions on Pattern Analysis and Machine Intelligence (2013)
9. Béthune, L., Vigouroux, D., Du, Y., VanRullen, R., Serre, T., Boutin, V.: Follow the Energy, Find the Path: Riemannian Metrics from Energy-Based Models. arXiv (2025)
10. Bodin, E., Stere, A.I., Margineantu, D.D., Ek, C.H., Moss, H.: Linear Combinations of Latents in Generative Models: Subspaces and Beyond. In: International Conference on Learning Representations (ICLR) (2025)
11. Charpiat, G., Girard, N., Felardos, L., Tarabalka, Y.: Input Similarity from the Neural Network Perspective. In: Advances in Neural Information Processing Systems (NeurIPS) (2019)
12. Chefer, H., Alaluf, Y., Vinker, Y., Wolf, L., Cohen-Or, D.: Attend-and-Excite: Attention-Based Semantic Guidance for Text-to-Image Diffusion Models. In: ACM Transactions on Graphics (TOG) (2023)
13. Chen, N., Klushyn, A., Kurle, R., Jiang, X., Bayer, J., Smagt, P.: Metrics for Deep Generative Models. In: International Conference on Artificial Intelligence and Statistics (AISTATS) (2018)
14. Chen, Z., Jiang, R., Duke, B., Zhao, H., Aarabi, P.: Exploring Gradient-Based Multi-directional Controls in GANs. In: European Conference on Computer Vision (ECCV) (2022)
15. Choi, J., Lee, J., Yoon, C., Park, J.H., Hwang, G., Kang, M.: Do Not Escape From the Manifold: Discovering the Local Coordinates on the Latent Space of GANs. In: International Conference on Learning Representations (ICLR) (2022)
16. Choi, Y., Uh, Y., Yoo, J., Ha, J.W.: StarGAN v2: Diverse Image Synthesis for Multiple Domains. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2020)
17. Chung, H., Kim, J., Park, G.Y., Nam, H., Ye, J.C.: CFG++: Manifold-constrained Classifier Free Guidance for Diffusion Models. In: International Conference on Learning Representations (ICLR) (2025)
18. Diepeveen, W., Batzolis, G., Shumaylov, Z., Schönlieb, C.B.: Score-based pullback riemannian geometry: Extracting the data manifold geometry using anisotropic flows. In: International Conference on Machine Learning (ICML) (2025)
19. Elfwing, S., Uchibe, E., Doya, K.: Sigmoid-Weighted Linear Units for Neural Network Function Approximation in Reinforcement Learning. arXiv (2017)
20. Fefferman, C., Mitter, S.K., Narayanan, H.: Testing the manifold hypothesis. Journal of the American Mathematical Society (2016)

21. Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A.H., Chechik, G., Cohen-Or, D.: An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion. In: International Conference on Learning Representations (ICLR) (2023)
22. George, A.J., Veiga, R., Macris, N.: Analysis of Diffusion Models for Manifold Data. arXiv (2025)
23. Goetschalckx, L., Andonian, A., Oliva, A., Isola, P.: GANalyze: Toward Visual Definitions of Cognitive Image Properties. In: IEEE/CVF International Conference on Computer Vision (ICCV) (2019)
24. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative Adversarial Nets. In: Advances in Neural Information Processing Systems (NeurIPS) (2014)
25. Gruffaz, S., Sassen, J.: Riemannian Metric Learning: Closer to You than You Imagine. arXiv (2025)
26. Guo, J., Xu, X., Pu, Y., Ni, Z., Wang, C., Vasu, M., Song, S., Huang, G., Shi, H.: Smooth Diffusion: Crafting Smooth Latent Spaces in Diffusion Models. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2024)
27. Haas, R., Graßhof, S., Brandt, S.S.: Tensor-based Emotion Editing in the StyleGAN Latent Space. In: CVPR 2022 Workshop on AI for Content Creation Workshop (2022)
28. Hahm, J., Lee, J., Kim, S., Lee, J.: Isometric representation learning for disentangled latent space of diffusion models. In: International Conference on Machine Learning (ICML) (2024)
29. Hanawa, K., Yokoi, S., Hara, S., Inui, K.: Evaluation of Similarity-based Explanations. In: International Conference on Learning Representations (ICLR) (2021)
30. Härkönen, E., Hertzmann, A., Lehtinen, J., Paris, S.: GANSpace: Discovering Interpretable GAN Controls. In: Advances in Neural Information Processing Systems (NeurIPS) (2020)
31. He, Q., Wang, J., Liu, Z., Yao, A.: AID: Attention Interpolation of Text-to-Image Diffusion. In: Advances in Neural Information Processing Systems (NeurIPS) (2024)
32. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In: Advances in Neural Information Processing Systems (NeurIPS) (2017)
33. Ho, J., Jain, A., Abbeel, P.: Denoising Diffusion Probabilistic Models. In: Advances in Neural Information Processing Systems (NeurIPS) (2020)
34. Ho, J., Salimans, T.: Classifier-Free Diffusion Guidance. In: NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications (2021)
35. Horvat, C., Pfister, J.P.: On Gauge Freedom, Conservativity and Intrinsic Dimensionality Estimation in Diffusion models. In: International Conference on Learning Representations (ICLR) (2024)
36. Humayun, A.I., Amara, I., Vasconcelos, C.N., Ramachandran, D., Schumann, C., He, J., Heller, K.A., Farnadi, G., Rostamzadeh, N., Havaei, M.: What Secrets Do Your Manifolds Hold? Understanding the Local Geometry of Generative Models. In: International Conference on Learning Representations (ICLR) (2025)
37. Jahanian, A., Chai, L., Isola, P.: On the "steerability" of generative adversarial networks. In: International Conference on Learning Representations, (ICLR) (2020)

38. Kamkari, H., Ross, B.L., Hosseinzadeh, R., Cresswell, J.C., Loaiza-Ganem, G.: A Geometric View of Data Complexity: Efficient Local Intrinsic Dimension Estimation with Diffusion Models. In: Advances in Neural Information Processing Systems (NeurIPS) (2024)
39. Karczewski, R., Heinonen, M., Garg, V.K.: Devil is in the Details: Density Guidance for Detail-Aware Generation with Flow Models. In: International Conference on Machine Learning (ICML) (2025)
40. Karczewski, R., Heinonen, M., Pouplin, A., Hauberg, S., Garg, V.: The Spacetime of Diffusion Models: An Information Geometry Perspective. In: International Conference on Learning Representations (ICLR) (2026)
41. Karras, T., Aila, T., Laine, S., Lehtinen, J.: Progressive Growing of GANs for Improved Quality, Stability, and Variation. In: International Conference on Learning Representations (ICLR) (2018)
42. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
43. Khrulkov, V., Mirvakhabova, L., Oseledets, I., Babenko, A.: Latent Transformations via NeuralODEs for GAN-based Image Editing. In: IEEE/CVF International Conference on Computer Vision (ICCV) (2021)
44. Kim, Y., Lee, K., Park, M., Na, B., chul Moon, I.: Diffusion Bridge AutoEncoders for Unsupervised Representation Learning. In: International Conference on Learning Representations (ICLR) (2025)
45. Kingma, D.P., Ba, J.: Adam: A Method for Stochastic Optimization. In: International Conference on Learning Representations (ICLR) (2015)
46. Kingma, D.P., Welling, M.: Auto-Encoding Variational Bayes. In: International Conference on Learning Representations (ICLR) (2014)
47. Kwon, M., Jeong, J., Uh, Y.: Diffusion Models Already Have A Semantic Latent Space. In: International Conference on Learning Representations (ICLR) (2023)
48. Kwon, M., seong Kim, S., Jeong, J., Hsiao, Y.T., Uh, Y.: TCFG: Tangential Damping Classifier-free Guidance. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2025)
49. Lee, J.M.: Introduction to Riemannian Manifolds. Springer (2019)
50. Lee, Y., Kim, S., Choi, J., Park, F.: A Statistical Manifold Framework for Point Cloud Data. In: International Conference on Machine Learning (ICML) (2022)
51. Li, J., Li, D., Savarese, S., Hoi, S.: BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In: International Conference on Machine Learning (ICML) (2023)
52. Liang, H., Hou, X., Shen, L.: SSFlow: Style-guided Neural Spline Flows for Face Image Manipulation. In: ACM International Conference on Multimedia (2021)
53. Lin, T.Y., Maire, M., Belongie, S.J., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: Common Objects in Context. In: European Conference on Computer Vision (ECCV) (2014)
54. Loaiza-Ganem, G., Ross, B.L., Hosseinzadeh, R., Caterini, A.L., Cresswell, J.C.: Deep Generative Models through the Lens of the Manifold Hypothesis: A Survey and New Connections. Transactions on Machine Learning Research (2024)
55. Lobashev, A., Guskov, D., Larchenko, M., Tamm, M.: Hessian Geometry of Latent Space in Generative Models. In: International Conference on Machine Learning (ICML) (2025)
56. Loshchilov, I., Hutter, F.: SGDR: Stochastic Gradient Descent with Warm Restarts. In: International Conference on Learning Representations (ICLR) (2017)

57. Loshchilov, I., Hutter, F.: Decoupled Weight Decay Regularization. In: International Conference on Learning Representations (ICLR) (2019)
58. Lu, Z., Wu, C., Chen, X., Wang, Y., Bai, L., Qiao, Y., Liu, X.: Hierarchical Diffusion Autoencoders and Disentangled Image Manipulation. In: IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) (2024)
59. Luo, S., Tan, Y., Huang, L., Li, J., Zhao, H.: Latent consistency models: Synthesizing high-resolution images with few-step inference. arXiv (2023)
60. Mokady, R., Hertz, A., Aberman, K., Pritch, Y., Cohen-Or, D.: Null-text Inversion for Editing Real Images using Guided Diffusion Models. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023 (2023)
61. Oldfield, J., Georgopoulos, M., Panagakis, Y., Nicolaou, M.A., Patras, I.: Tensor Component Analysis for Interpreting the Latent Space of GANs. In: British Machine Vision Conference (BMVC) (2021)
62. Park, Y.H., Kwon, M., Choi, J., Jo, J., Uh, Y.: Understanding the Latent Space of Diffusion Models through the Lens of Riemannian Geometry. In: Advances in Neural Information Processing Systems (NeurIPS) (2023)
63. Park, Y.H., Kwon, M., Jo, J., Uh, Y.: Unsupervised Discovery of Semantic Latent Directions in Diffusion Models. arXiv (2023)
64. Perazzi, F., Pont-Tuset, J., McWilliams, B., Van Gool, L., Gross, M., Sorkine-Hornung, A.: A benchmark dataset and evaluation methodology for video object segmentation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
65. Pidstrigach, J.: Score-Based Generative Models Detect Manifolds. In: Advances in Neural Information Processing Systems (NeurIPS) (2022)
66. von Platen, P., Patil, S., Lozhkov, A., Cuenca, P., Lambert, N., Rasul, K., Davaadorj, M., Nair, D., Paul, S., Berman, W., Xu, Y., Liu, S., Wolf, T.: Diffusers: State-of-the-art diffusion models. `https://github.com/huggingface/diffusers` (2022), last accessed 15 Feb 2026
67. Plumerault, A., Borgne, H.L., Hudelot, C.: Controlling generative models with continuous factors of variations. In: International Conference on Learning Representations (ICLR) (2020)
68. Potaptchik, P., Azangulov, I., Deligiannidis, G.: Linear Convergence of Diffusion Models Under the Manifold Hypothesis. arXiv (2025)
69. Preechakul, K., Chatthee, N., Wizadwongsa, S., Suwajanakorn, S.: Diffusion Autoencoders: Toward a Meaningful and Decodable Representation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2022)
70. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning (ICML) (2021)
71. Ramesh, A., Choi, Y., LeCun, Y.: A Spectral Regularizer for Unsupervised Disentanglement. arXiv (2019)
72. Rassin, R., Hirsch, E., Glickman, D., Ravfogel, S., Goldberg, Y., Chechik, G.: Linguistic Binding in Diffusion Models: Enhancing Attribute Correspondence through Attention Map Alignment. In: Advances in Neural Information Processing Systems (NeurIPS) (2023)
73. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution Image Synthesis with Latent Diffusion Models. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2022)

74. Samuel, D., Ben-Ari, R., Darshan, N., Maron, H., Chechik, G.: Norm-guided Latent Space Exploration for Text-to-image Generation. In: Advances in Neural Information Processing Systems (NeurIPS) (2023)
75. Shao, H., Kumar, A., Fletcher, P.T.: The Riemannian Geometry of Deep Generative Models. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) (2018)
76. Shen, L., Liu, T., Sun, H., Ye, X., Li, B., Zhang, J., Cao, Z.: DreamMover: Leveraging the Prior of Diffusion Models for Image Interpolation with Large Motion. In: European Conference on Computer Vision (ECCV) (2024)
77. Shen, Y., Gu, J., Tang, X., Zhou, B.: Interpreting the Latent Space of GANs for Semantic Face Editing. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2020)
78. Shen, Y., Zhou, B.: Closed-form Factorization of Latent Semantics in GANs. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2021)
79. Shoemake, K.: Animating Rotation with Quaternion Curves. International Conference on Computer Graphics and Interactive Techniques (1985)
80. Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S.: Deep Unsupervised Learning using Nonequilibrium Thermodynamics. In: International Conference on Machine Learning (ICML) (2015)
81. Song, J., Meng, C., Ermon, S.: Denoising Diffusion Implicit Models. In: International Conference on Learning Representations (ICLR) (2021)
82. Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-Based Generative Modeling through Stochastic Differential Equations. In: International Conference on Learning Representations (ICLR) (2021)
83. Sorrenson, P., Behrend-Uriarte, D., Schnörr, C., Köthe, U.: Learning distances from data with normalizing flows and score matching. In: International Conference on Machine Learning (ICML) (2025)
84. Spingarn, N., Banner, R., Michaeli, T.: GAN ”steerability” without optimization. In: International Conference on Learning Representations (ICLR) (2021)
85. Stanczuk, J.P., Batzolis, G., Deveney, T., Schönlieb, C.B.: Diffusion Models Encode the Intrinsic Dimension of Data Manifolds. In: International Conference on Machine Learning (ICML) (2024)
86. Sueyoshi, K., Matsubara, T.: Predicated Diffusion: Predicate Logic-Based Attention Guidance for Text-to-Image Diffusion Models. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2024)
87. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the Inception Architecture for Computer Vision. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
88. Tang, R., Yang, Y.: Adaptivity of Diffusion Models to Manifold Structures. In: International Conference on Artificial Intelligence and Statistics (AISTATS) (2024)
89. Tewari, A.K., Elgharib, M.A., Bharaj, G., Bernard, F., Seidel, H.P., Pérez, P., Zollhöfer, M., Theobalt, C.: StyleRig: Rigging StyleGAN for 3D Control Over Portrait Images. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2020)
90. Tzelepis, C., Tzimiropoulos, G., Patras, I.: WarpedGANSpace: Finding non-linear RBF paths in GAN latent space. In: IEEE/CVF International Conference on Computer Vision (ICCV) (2021)
91. Ventura, E., Achilli, B., Silvestri, G., Lucibello, C., Ambrogioni, L.: Manifolds, Random Matrices and Spectral Gaps: The Geometric Phases of Generative Diffusion. In: International Conference on Learning Representations (ICLR) (2025)

92. Voynov, A., Babenko, A.: Unsupervised discovery of interpretable directions in the GAN latent space. In: International Conference on Machine Learning (ICML) (2020)
93. Wang, C.J., Golland, P.: Interpolating between Images with Diffusion Models. In: ICML 2023 Workshop on Challenges of Deploying Generative AI (2023)
94. Wenliang, L.K., Moran, B.: Score-based generative models learn manifold-like structures with constrained mixing. In: NeurIPS 2022 Workshop on Score-Based Methods (2022)
95. Yang, T., Arvanitidis, G., Fu, D., Li, X., Hauberg, S.: Geodesic Clustering in Deep Generative Models. arXiv (2018)
96. Yang, Z., Yu, Z., Xu, Z., Singh, J., Zhang, J., Campbell, D., Tu, P., Hartley, R.: IMPUS: Image Morphing with Perceptually-Uniform Sampling Using Diffusion Models. In: International Conference on Learning Representations (ICLR) (2024)
97. Yeh, C.K., Kim, J.S., Yen, I.E.H., Ravikumar, P.: Representer Point Selection for Explaining Deep Neural Networks. In: Advances in Neural Information Processing Systems (NeurIPS) (2018)
98. Yu, Q., Singh, J., Yang, Z., Tu, P.H., Zhang, J., Li, H., Hartley, R., Campbell, D.: Probability Density Geodesics in Image Diffusion Latent Space. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2025)
99. Yun, Z., Chuang, G., Dong, D., Chen, Y.: Denoising for Manifold Extrapolation. In: NeurIPS 2024 Workshop on Scientific Methods for Understanding Deep Learning (SciForDL) (2024)
100. Zhang, K., Zhou, Y., Xu, X., Dai, B., Pan, X.: DiffMorpher: Unleashing the Capability of Diffusion Models for Image Morphing. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2024)
101. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2018)
102. Zheng, P., Zhang, Y., Fang, Z., Liu, T., Lian, D., Han, B.: NoiseDiffusion: Correcting Noise for Image Interpolation with Diffusion Models beyond Spherical Linear Interpolation. In: International Conference on Learning Representations (ICLR) (2024)
103. Zhou, T., Tucker, R., Flynn, J., Fyffe, G., Snavely, N.: Stereo magnification: Learning view synthesis using multiplane images. ACM Transactions on Graphics (TOG) (2018)
104. Zhu, T., Ren, D., Wang, Q., Wu, X., Zuo, W.: Generative inbetweening through frame-wise conditions-driven video generation. arXiv (2024)
105. Zhuang, P., Koyejo, O.O., Schwing, A.: Enjoy Your Editing: Controllable GANs for image editing via latent space navigation. In: International Conference on Learning Representations (ICLR) (2021)