Extracting the Data Manifold from Diffusion Models via a Score-Based Non-Conformal Riemannian Metric

Shinnosuke Saito [1] **Takashi Matsubara** [2 1]

Abstract

Diffusion models generate high-quality samples, but unlike VAEs or GANs they do not provide an explicit low-dimensional latent space that parameterizes the data manifold. This makes manifold-aware operations, such as geometrically faithful interpolation and guidance that stays on the learned manifold, difficult to formulate. We propose a training-free Riemannian metric on the noise space derived from the score Jacobian. The metric exploits its spectral structure, which separates tangent and normal directions of the data manifold, and therefore encourages paths and guidance updates to remain tangential to the manifold. It gives a single geometric tool for geodesic interpolation and correction of classifier-free guidance. Experiments on synthetic data, image and video interpolation, and text-to-image guidance show that the proposed metric preserves manifold geometry better than density-based alternatives.

1. Introduction

Diffusion models are a class of deep generative models that can generate diverse, high-fidelity samples (Ho et al., 2020; Song et al., 2021a; Rombach et al., 2022). Their behavior is often understood through the manifold hypothesis: natural data concentrate near a low-dimensional manifold embedded in a high-dimensional ambient space (Bengio et al., 2012; Fefferman et al., 2016; Loaiza-Ganem et al., 2024). For VAEs and GANs, this viewpoint leads naturally to Riemannian geometry. A decoder or generator maps a low-dimensional latent space to data space, and pulling back the Euclidean metric through this map yields geodesics that respect the learned geometry (Shao et al., 2017; Arvanitidis et al., 2018; Chen et al., 2018).

[1]Hokkaido University, Sapporo, Japan [2]CyberAgent AI Lab, Tokyo, Japan. Correspondence to: Shinnosuke Saito <saitou.shinnosuke.y0@elms.hokudai.ac.jp>, Takashi Matsubara <matsubara@ist.hokudai.ac.jp>.

Published as a paper at the 1st FoGen workshop, ICML 2026, Seoul, South Korea, 2026. Copyright 2026 by the author(s).

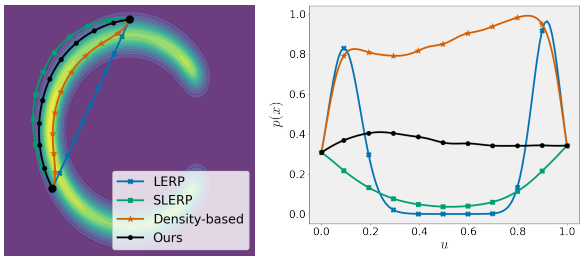

Figure 1. **Synthetic example.** On a C-shaped distribution, LERP crosses low-density regions, SLERP deviates from the manifold, and density-based geodesics drift to high-density regions. Ours runs parallel to the manifold and preserves the endpoint densities.

Diffusion models lack such a parameterization, making decoder-based pullback metrics unavailable. Recent methods define metrics from the density of noisy samples or from directions toward the data manifold (Yu et al., 2025; Azeglio & Bernardo, 2025). However, these cues indicate where samples concentrate or where the manifold lies, but not how it is locally oriented; moreover, high likelihood need not coincide with perceptual detail. Consequently, density-based geodesics can be pulled toward high-density regions and produce over-smoothed images (Karczewski et al., 2025). This motivates a metric that reflects local orientation rather than density alone.

Prior work shows that the score function points roughly normal to the data manifold and that its Jacobian exhibits a spectral gap separating normal and tangent directions (Stanczuk et al., 2024; Ventura et al., 2025). Based on this observation, we define a Riemannian metric whose matrix is $J_{x_t}^\top J_{x_t}$. The resulting metric assigns large cost to normal movement and small cost to tangent movement. This provides a geometry for diffusion noise spaces without training additional networks or changing the diffusion model.

Our contribution is a score-Jacobian metric and its use as a tangent-normal lens for diffusion models. We validate this perspective from two complementary angles. Globally, geodesics under the metric yield natural image interpolations while avoiding the over-smoothing of density-based geodesics. Locally, the same metric can correct classifier-free guidance (CFG) by suppressing guidance components that move samples away from the learned manifold.

2. Related Work

Riemannian geometry of generative models. Riemannian metrics have been widely used to analyze and manipulate latent spaces of VAEs and GANs. Some methods learn auxiliary metrics (Yang et al., 2018; Arvanitidis et al., 2022; Lee et al., 2022; Sorrenson et al., 2025), while others construct training-free pullback geometries from pre-trained generative maps (Shao et al., 2017; Chen et al., 2018; Arvanitidis et al., 2018; 2021). A closely related score-based construction focuses on unimodal distributions and learns an anisotropic normalizing flow to induce a pullback geometry and global chart (Diepeveen et al., 2025), whereas we directly define a training-free metric on the noise space from the Jacobian spectrum of a pre-trained score function.

Manifold and interpolation in diffusion models. Diffusion models implicitly learn manifold structure through denoising (Pidstrigach, 2022; Wenliang & Moran, 2022; Tang & Yang, 2024; Yun et al., 2024; Potaptchik et al., 2025). In particular, the score Jacobian has been used to estimate local intrinsic dimension and identify tangent and normal directions (Stanczuk et al., 2024; Ventura et al., 2025). For interpolation, closed-form methods such as LERP and SLERP ignore this local geometry (Ho et al., 2020; Shoemake, 1985; Song et al., 2021a), while training-dependent methods modify the model or introduce additional networks (Zhang et al., 2023; Guo et al., 2024; Hahm et al., 2024). The closest line of work is training-free Riemannian geometry, including density geodesics and FIM-inspired score metrics (Yu et al., 2025; Azeglio & Bernardo, 2025). Our metric also follows this training-free geometric view, but uses the full score-Jacobian spectrum to penalize all local normal directions rather than relying on density or rank-one score information.

Classifier-free guidance. Text-to-image diffusion models condition the denoising process on a prompt, and classifier-free guidance (CFG) amplifies this conditioning to improve prompt fidelity (Ho & Salimans, 2021; Rombach et al., 2022). Negative prompts and strong guidance can also shape the trajectory, but they may distort the learned data manifold and produce unnatural images (Rombach et al., 2022). Recent methods mitigate this distortion: CFG++ uses the unconditional score for renoising, and TCFG projects the unconditional score onto a shared conditional subspace (Chung et al., 2025; Kwon et al., 2025). These corrections are motivated by manifold preservation, but they do not formally characterize the geometry; our metric instead decomposes guidance into tangent and normal components.

3. Method

3.1. Score-Jacobian Metric

Let $x_t \in \mathbb{R}^D$ be a noisy sample at diffusion time t, let $s_\theta(x_t, t)$ be the score function of a pre-trained diffusion model, and let $v, w \in T_{x_t}\mathbb{R}^D$ be tangent vectors in the ambient noise space. We define the metric

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

where $J_{x_t} = \nabla_{x_t} s_\theta(x_t, t)$. Equivalently, the metric measures the Euclidean change of the score along each direction. At the intermediate noise levels used below, G_{x_t} is full rank in practice and its spectrum defines a tangent-normal split.

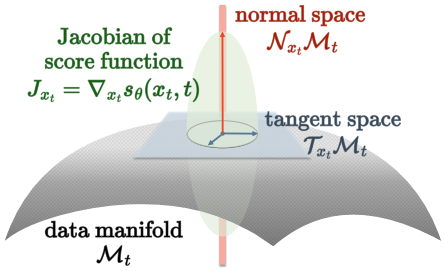

Figure 2. **Metric construction.** The score Jacobian separates tangent and normal directions of the learned manifold. The proposed metric penalizes movement in normal directions.

The interpretation follows from the spectral structure of J_{x_t}. Let $\mathcal{T}_{x_t}\mathcal{M}_t$ be the subspace spanned by right singular vectors associated with small singular values, and let $\mathcal{N}_{x_t}\mathcal{M}_t$ be its orthogonal complement. These spaces approximate tangent and normal spaces of the learned data manifold $\mathcal{M}_t$ (Stanczuk et al., 2024; Ventura et al., 2025). For a fixed Euclidean norm, $g_{x_t}(v, v) = \|J_{x_t} v\|_2^2$ is small along tangent directions and large along normal directions.

Proposition 1. *Among ambient directions $v \in T_{x_t}\mathbb{R}^D$ of 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 $\mathcal{T}_{x_t}\mathcal{M}_t$ induced by the small-singular-value subspace of J_{x_t}.*

This construction differs from density-based geodesics, which can be attracted toward high-density modes, because it penalizes changes of the score field and therefore emphasizes local manifold orientation. It also differs from the FIM-inspired metric $\lambda s_\theta s_\theta^\top + I$: the score outer product captures only one normal direction, whereas $J_{x_t}^\top J_{x_t}$ uses the full local spectrum and can penalize multiple normal directions in high-codimension image manifolds.

Thus, short paths under our metric are encouraged to stay on or parallel to the data manifold. Figure 2 illustrates the resulting tangent-normal decomposition; see Appendix B.

3.2. Geodesic Interpolation

For a curve $\gamma : [0, 1] \to \mathbb{R}^D$, the energy under Equation (1) is

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

Table 1. **Results on synthetic 2D data.**

Method	Std. ↓
LERP	0.1606
SLERP	0.0833
Density	0.1073
Ours	**0.0701**

The first expression shows that geodesics minimize normal movement. The second, obtained by the chain rule, shows that they minimize score variation along the path. Since scores are gradients of log densities, minimizing their variation keeps the path in a consistent relation to the density landscape rather than explicitly seeking high-density modes. Once a geodesic is obtained, the endpoint distance is $d_g(\gamma(0), \gamma(1)) = \int_0^1 \sqrt{g_{\gamma(u)}(\gamma'(u), \gamma'(u))}\, \mathrm{d}u$.

In practice, we interpolate at a fixed time $\tau > 0$. Given endpoints $x_0^{(0)}$ and $x_0^{(1)}$, we map them to $x_\tau^{(0)}$ and $x_\tau^{(1)}$ using DDIM inversion. We discretize the path into $N+1$ points and minimize

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

over intermediate points, initialized by SLERP. The corresponding discrete distance is $d_g \approx \sum_{i=0}^{N-1} \|s_\theta(x_\tau^{(u_{i+1})}, \tau) - s_\theta(x_\tau^{(u_i)}, \tau)\|_2$. The optimized noisy points are then mapped back to clean samples by deterministic denoising; see Algorithm 1.

3.3. Guidance Correction

The same metric gives a local correction for CFG. For the diffusion update $x_{t-1} = x_t + \eta_t s_\theta(x_t, t, c)$, CFG adds a guidance term $\Delta s(x_t, t, c)$. We seek a corrected guidance $\Delta\hat{s}$ whose sample remains close to the CFG-guided sample in Euclidean distance while staying close to the unguided sample under the proposed metric:

$$\begin{aligned}\Delta\hat{s}^* = \arg\min_{\Delta\hat{s}}\ & d_E(x_t + \eta_t(s_\theta + \Delta s), x_t + \eta_t(s_\theta + \Delta\hat{s}))^2 \\ & + \lambda d_g(x_t + \eta_t s_\theta,\ x_t + \eta_t(s_\theta + \Delta\hat{s}))^2.\end{aligned} \quad (4)$$

For tractability, we evaluate d_g at t, approximate it with one local segment, and absorb η_t into λ; this reduces Equation (4) to the linear system $(I + \lambda G_{x_t})\Delta\hat{s}^* = \Delta s$. We solve it with one conjugate-gradient step initialized at Δs:

$$\begin{aligned}\Delta\hat{s}^* &= \Delta s + \frac{\epsilon^\top \epsilon}{\epsilon^\top (I + \lambda G_{x_t})\epsilon}\epsilon \\ \text{for} \quad \epsilon &= \Delta s - (I + \lambda G_{x_t})\Delta s.\end{aligned} \quad (5)$$

Multiplication by $G_{x_t} = J_{x_t}^\top J_{x_t}$ is approximated with finite-difference Jacobian-vector products, requiring four additional score evaluations per denoising step; see Algorithm 2.

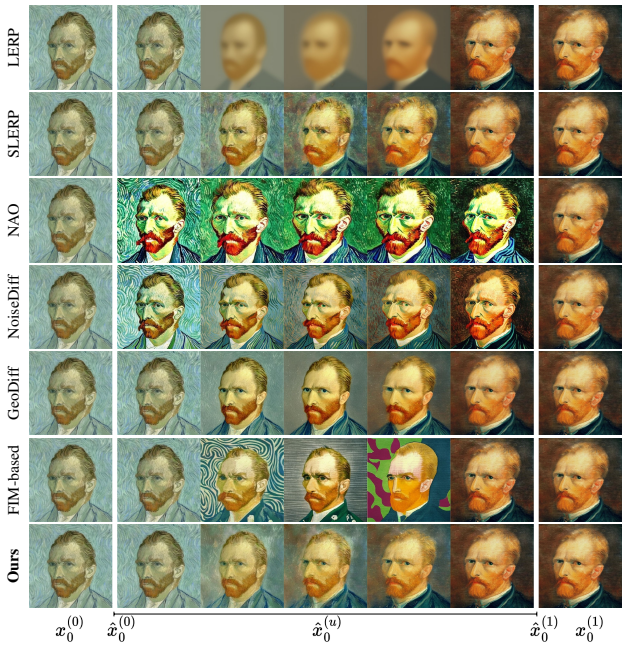

Figure 3. **Qualitative image interpolation.** Ours keeps endpoint details through the path, while density- or norm-based baselines often over-smooth the sequence or alter endpoint appearance.

4. Experiments

4.1. Synthetic 2D Data

We train a DDPM on a C-shaped 2D distribution with $T = 50$ and interpolate at $\tau = 1$. As shown in Figure 1, our geodesic runs parallel to the manifold rather than crossing low-density regions or moving toward high-density modes. As reported in Table 1, our metric achieves the lowest standard deviation of density along the path over 50 endpoint pairs; see Appendix D.1.

4.2. Image Interpolation

We use Stable Diffusion v2.1-base (Rombach et al., 2022) with $T = 50$, $\tau = 0.6T$, and $N - 1 = 9$ interpolated images. We evaluate on MorphBench Animation and Metamorphosis (Zhang et al., 2023), CelebA-HQ (Karras et al., 2018), and Animal Faces-HQ (Choi et al., 2020). Baselines are LERP, SLERP, NAO, NoiseDiffusion, GeodesicDiffusion, and FIM-based metric (Ho et al., 2020; Shoemake, 1985; Song et al., 2021a; Samuel et al., 2023; Zheng et al., 2024; Yu et al., 2025; Azeglio & Bernardo, 2025); see Appendices C.1 and D.2. We report the full quantitative results in Table 2. Our metric achieves the best FID, PPL, and endpoint reconstruction error on all datasets, and records the best or second-best PDV. Qualitatively, LERP blurs, norm-adjusting baselines alter endpoints, and density-based and FIM-based methods over-smooth or cartoonize texture; ours preserves endpoint details throughout the path (Figure 3).

Table 2. **Image interpolation results.**

Method	FID ↓				PPL ↓				PDV ↓				RE ↓ ($\times 10^{-3}$)			
	MB(A)	MB(M)	CA	AF	MB(A)	MB(M)	CA	AF	MB(A)	MB(M)	CA	AF	MB(A)	MB(M)	CA	AF
LERP	84.20	118.90	95.68	119.58	0.848	1.787	1.420	1.859	0.055	0.128	0.091	0.154	0.401	0.397	1.010	2.049
SLERP	62.81	48.99	37.84	26.07	0.644	1.065	0.707	0.871	0.030	**0.055****	**0.033***	**0.022***	0.401	0.397	1.010	2.049
NAO	130.54	102.64	83.05	71.47	2.868	4.299	2.121	2.443	0.163	0.164	0.154	0.173	39.244	44.302	27.623	40.178
NoiseDiff	119.47	74.03	65.04	68.87	3.618	2.011	2.098	3.250	0.064	0.085	0.069	0.083	15.096	7.835	8.618	19.628
GeoDiff	28.70	38.12	35.98	25.80	0.402	1.021	0.669	0.842	0.024	0.073	0.044	0.027	0.188	0.272	0.891	1.969
FIM-based	92.09	78.80	70.95	59.11	3.358	4.429	4.152	5.249	0.142	0.187	0.172	0.196	0.401	0.397	1.010	2.049
Ours	**27.44**	**36.00**	**32.54**	**21.01**	**0.380***	**0.977****	**0.633****	**0.767****	**0.021***	0.073	0.036	0.023	**0.177***	**0.201****	**0.888****	**1.962****

Bold indicates the best result and underline indicates the second best. * 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 ($H_0 : p = 0.5$).

Table 3. **Video frame interpolation results.**

Method	MSE ↓ ($\times 10^{-3}$)			LPIPS ↓		
	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****

Table 4. **Guidance correction results.**

Method	$w = 5.0$		$w = 7.5$		$w = 12.5$	
	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

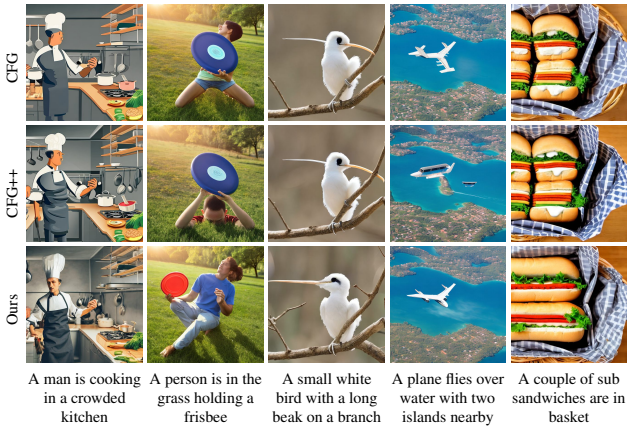

Figure 4. **Qualitative guidance comparison.** Ours suppresses off-manifold artifacts while preserving the conditioning signal. Prompts are shown below the images.

4.3. Video Frame Interpolation

Image metrics are indirect, so we also compare against ground-truth middle frames. Following prior protocol (Zhu et al., 2024), we use three-frame clips from DAVIS (Perazzi et al., 2016), Pexels Human, and RealEstate10K (Zhou et al., 2018); the first and third frames are endpoints and the second frame is the target. All methods use the same settings as image interpolation; see Appendix E.1. Table 3 shows that our method achieves the best MSE and LPIPS on all datasets, indicating interpolations closer to true middle frames.

4.4. Guidance Correction

We evaluate 30K MS-COCO 2014 captions (Lin et al., 2014) with Stable Diffusion v2.1-base using FID and CLIP Score (Heusel et al., 2017; Radford et al., 2021). Table 4 shows lower FID than CFG and CFG++ at every guidance scale without degrading CLIP Score. Qualitatively, the correction reduces prompt over-emphasis and cartoon-like textures (Figure 4). For 4-step LCM distillation, applying the correction on the teacher improves FID from 17.91 to 16.98 while preserving CLIP Score at 0.306, with no inference overhead. See Appendices D.3 and E.2 for details.

5. Conclusion

We proposed a training-free Riemannian metric for diffusion models built from the score Jacobian. By exploiting its spectral structure, the metric separates tangent and normal directions of the data manifold and penalizes off-manifold motion. We used the same metric for two complementary operations: global geodesic interpolation and local CFG correction. Experiments on synthetic data, image and video interpolation, and text-to-image guidance show that this geometry preserves endpoint details and suppresses off-manifold artifacts. Future work includes faster Jacobian-vector approximations and extensions to flow matching and broader manifold-aware editing tasks.

Impact Statement

This paper presents a geometric analysis and control method for diffusion models. It is not expected to have any direct negative impact on society or individuals beyond the general risks associated with generative models.

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

Arvanitidis, G., Hansen, L. K., and Hauberg, S. Latent Space Oddity: on the Curvature of Deep Generative Models. In *International Conference on Learning Representations (ICLR)*, 2018.

Arvanitidis, G., Hauberg, S., and Schölkopf, B. Geometrically Enriched Latent Spaces. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2021.

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

Azeglio, S. and Bernardo, A. D. What's Inside Your Diffusion Model? A Score-Based Riemannian Metric to Explore the Data Manifold. *arXiv*, 2025.

Bengio, Y., Courville, A. C., and Vincent, P. Representation Learning: A Review and New Perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2012.

Changpinyo, S., Sharma, P., Ding, N., and Soricut, R. Conceptual 12M: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

Chen, N., Klushyn, A., Kurle, R., Jiang, X., Bayer, J., and Smagt, P. Metrics for Deep Generative Models. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2018.

Choi, Y., Uh, Y., Yoo, J., and Ha, J.-W. StarGAN v2: Diverse Image Synthesis for Multiple Domains. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

Chung, H., Kim, J., Park, G. Y., Nam, H., and Ye, J. C. CFG++: Manifold-constrained Classifier Free Guidance for Diffusion Models. In *International Conference on Learning Representations (ICLR)*, 2025.

Diepeveen, W., Batzolis, G., Shumaylov, Z., and Schönlieb, C.-B. Score-based pullback riemannian geometry: Extracting the data manifold geometry using anisotropic flows. In *International Conference on Machine Learning (ICML)*, 2025.

Elfwing, S., Uchibe, E., and Doya, K. Sigmoid-Weighted Linear Units for Neural Network Function Approximation in Reinforcement Learning. *arXiv*, 2017.

Fefferman, C., Mitter, S. K., and Narayanan, H. Testing the manifold hypothesis. *Journal of the American Mathematical Society*, 2016.

Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A. H., Chechik, G., and 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.

Guo, J., Xu, X., Pu, Y., Ni, Z., Wang, C., Vasu, M., Song, S., Huang, G., and Shi, H. Smooth Diffusion: Crafting Smooth Latent Spaces in Diffusion Models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Hahm, J., Lee, J., Kim, S., and Lee, J. Isometric representation learning for disentangled latent space of diffusion models. In *International Conference on Machine Learning (ICML)*, 2024.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and 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.

Ho, J. and Salimans, T. Classifier-Free Diffusion Guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021.

Ho, J., Jain, A., and Abbeel, P. Denoising Diffusion Probabilistic Models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

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

Hu, E. J., yelong shen, Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022.

Kamkari, H., Ross, B. L., Hosseinzadeh, R., Cresswell, J. C., and 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.

Karczewski, R., Heinonen, M., and 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.

Karras, T., Aila, T., Laine, S., and Lehtinen, J. Progressive Growing of GANs for Improved Quality, Stability, and Variation. In *International Conference on Learning Representations (ICLR)*, 2018.

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

Kwon, M., seong Kim, S., Jeong, J., Hsiao, Y. T., and Uh, Y. TCFG: Tangential Damping Classifier-free Guidance. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

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

Lee, Y., Kim, S., Choi, J., and Park, F. A Statistical Manifold Framework for Point Cloud Data. In *International Conference on Machine Learning (ICML)*, 2022.

Li, J., Li, D., Xiong, C., and Hoi, S. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. In *International Conference on Machine Learning (ICML)*, 2022.

Lin, T.-Y., Maire, M., Belongie, S. J., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. Microsoft COCO: Common Objects in Context. In *European Conference on Computer Vision (ECCV)*, 2014.

Loaiza-Ganem, G., Ross, B. L., Hosseinzadeh, R., Caterini, A. L., and Cresswell, J. C. Deep Generative Models through the Lens of the Manifold Hypothesis: A Survey and New Connections. *Transactions on Machine Learning Research*, 2024.

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

Loshchilov, I. and Hutter, F. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations (ICLR)*, 2019.

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

Luo, S., Tan, Y., Huang, L., Li, J., and Zhao, H. Latent consistency models: Synthesizing high-resolution images with few-step inference. *arXiv*, 2023.

Mokady, R., Hertz, A., Aberman, K., Pritch, Y., and 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.

Perazzi, F., Pont-Tuset, J., McWilliams, B., Van Gool, L., Gross, M., and 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.

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

Potaptchik, P., Azangulov, I., and Deligiannidis, G. Linear Convergence of Diffusion Models Under the Manifold Hypothesis. *arXiv*, 2025.

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.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution Image Synthesis with Latent Diffusion Models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Samuel, D., Ben-Ari, R., Darshan, N., Maron, H., and Chechik, G. Norm-guided Latent Space Exploration for Text-to-image Generation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Shao, H., Kumar, A., and Fletcher, P. T. The Riemannian Geometry of Deep Generative Models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2017.

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

Song, J., Meng, C., and Ermon, S. Denoising Diffusion Implicit Models. In *International Conference on Learning Representations (ICLR)*, 2021a.

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. Score-Based Generative Modeling through Stochastic Differential Equations. In *International Conference on Learning Representations (ICLR)*, 2021b.

Sorrenson, P., Behrend-Uriarte, D., Schnörr, C., and Köthe, U. Learning distances from data with normalizing flows and score matching. In *International Conference on Machine Learning (ICML)*, 2025.

Stanczuk, J. P., Batzolis, G., Deveney, T., and Schönlieb, C.-B. Diffusion Models Encode the Intrinsic Dimension of Data Manifolds. In *International Conference on Machine Learning (ICML)*, 2024.

Tang, R. and Yang, Y. Adaptivity of Diffusion Models to Manifold Structures. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2024.

Ventura, E., Achilli, B., Silvestri, G., Lucibello, C., and Ambrogioni, L. Manifolds, Random Matrices and Spectral Gaps: The Geometric Phases of Generative Diffusion. In *International Conference on Learning Representations (ICLR)*, 2025.

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., and Wolf, T. Diffusers: State-of-the-art diffusion models. https://github.com/huggingface/diffusers, 2022.

Wenliang, L. K. and Moran, B. Score-based generative model learn manifold-like structures with constrained mixing. In *NeurIPS 2022 Workshop on Score-Based Methods*, 2022.

Yang, T., Arvanitidis, G., Fu, D., Li, X., and Hauberg, S. Geodesic Clustering in Deep Generative Models. *arXiv*, 2018.

Yu, Q., Singh, J., Yang, Z., Tu, P. H., Zhang, J., Li, H., Hartley, R., and Campbell, D. Probability Density Geodesics in Image Diffusion Latent Space. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Yun, Z., Chuang, G., Dong, D., and Chen, Y. Denoising for Manifold Extrapolation. In *NeurIPS 2024 Workshop on Scientific Methods for Understanding Deep Learning (SciForDL)*, 2024.

Zhang, K., Zhou, Y., Xu, X., Pan, X., and Dai, B. DiffMorpher: Unleashing the Capability of Diffusion Models for Image Morphing. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Zheng, P., Zhang, Y., Fang, Z., Liu, T., Lian, D., and Han, B. NoiseDiffusion: Correcting Noise for Image Interpolation with Diffusion Models beyond Spherical Linear Interpolation. In *International Conference on Learning Representations (ICLR)*, 2024.

Zhou, T., Tucker, R., Flynn, J., Fyffe, G., and Snavely, N. Stereo magnification: Learning view synthesis using multiplane images. *ACM Transactions on Graphics (TOG)*, 2018.

Zhu, T., Ren, D., Wang, Q., Wu, X., and Zuo, W. Generative inbetweening through frame-wise conditions-driven video generation. *arXiv*, 2024.

A. Preliminaries

A.1. Riemannian Geometry

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

$$g_p(v, w) = g_p(w, v), \quad g_p(v, v) \geq 0 \text{ for all } v \in T_p\mathcal{M}, \quad g_p(v, v) = 0 \Leftrightarrow v = 0. \tag{6}$$

By identifying g_p with an inner product, we write

$$\langle v, w \rangle_g \coloneqq g_p(v, w) \quad \text{for any } v, w \in T_p\mathcal{M}. \tag{7}$$

The pair $(\mathcal{M}, g)$ is called a Riemannian manifold.

Let $(x^1, \ldots, x^D)$ be smooth local coordinates in a neighborhood of $p \in \mathcal{M}$. The coordinate basis for $T_p\mathcal{M}$ is $\left(\frac{\partial}{\partial x^1}|_p, \ldots, \frac{\partial}{\partial x^D}|_p\right)$. Tangent vectors $v, w \in T_p\mathcal{M}$ can be expressed as $v = \sum_{i=1}^D v^i \frac{\partial}{\partial x^i}|_p$ and $w = \sum_{i=1}^D w^i \frac{\partial}{\partial x^i}|_p$. The matrix representation G_p of g at p has entries

$$g_{ij}(p) = g_p\left(\frac{\partial}{\partial x^i}\Big|_p, \frac{\partial}{\partial x^j}\Big|_p\right) = \left\langle \frac{\partial}{\partial x^i}\Big|_p, \frac{\partial}{\partial x^j}\Big|_p \right\rangle_g. \tag{8}$$

The matrix G_p is symmetric and positive definite, and the Euclidean metric corresponds to $G_p = I$. The inner product of two tangent vectors is then

$$g_p(v, w) = \sum_{i=1}^D \sum_{j=1}^D g_{ij}(p) v^i w^j = v^\top G_p w. \tag{9}$$

Geodesics. The length of a tangent vector $v \in T_p\mathcal{M}$ is $|v|_g \coloneqq \sqrt{\langle v, v \rangle_g}$. For a smooth curve $\gamma : [0, 1] \to \mathcal{M}$, $u \mapsto \gamma(u)$, its length is

$$L[\gamma] \coloneqq \int_0^1 |\gamma'(u)|_g \, \mathrm{d}u = \int_0^1 \sqrt{\langle \gamma'(u), \gamma'(u) \rangle_g} \, \mathrm{d}u = \int_0^1 \sqrt{\gamma'(u)^\top G_{\gamma(u)} \gamma'(u)} \, \mathrm{d}u. \tag{10}$$

A *geodesic* is a curve that locally minimizes length; intuitively, it is a locally shortest path between two points. It is often more convenient to work with the energy functional

$$E[\gamma] = \frac{1}{2} \int_0^1 |\gamma'(u)|_g^2 \, \mathrm{d}u = \frac{1}{2} \int_0^1 \langle \gamma'(u), \gamma'(u) \rangle_g \, \mathrm{d}u. \tag{11}$$

The critical points of this energy are constant-speed geodesics. Geodesics can also be obtained by solving a second-order geodesic equation, but doing so generally requires $O(D^3)$ computation and is infeasible in high-dimensional diffusion noise spaces. We therefore optimize the discretized energy in Equation (3).

A.2. Diffusion Models

Forward Process Let $x_0 \in \mathbb{R}^D$ be a data sample. The forward process is a Markov chain that recursively adds Gaussian noise at timesteps $t = 1, \ldots, T$:

$$q(x_t \mid x_{t-1}) = \mathcal{N}\left(x_t; \sqrt{1 - \beta_t} x_{t-1}, \beta_t I\right) = \mathcal{N}\left(x_t; \sqrt{\frac{\alpha_t}{\alpha_{t-1}}} x_{t-1}, \left(1 - \frac{\alpha_t}{\alpha_{t-1}}\right) I\right), \tag{12}$$

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

Denoising Process. The generation process of diffusion models is the denoising, or reverse, process. Starting from $x_T \sim \mathcal{N}(0, I)$, it iteratively removes noise from $t = T$ to $t = 0$ and obtains a clean sample x_0. A standard DDPM reverse Markov chain is written as

$$x_{t-1} = \frac{1}{\sqrt{1 - \beta_t}} \left(x_t - \frac{\beta_t}{\sqrt{1 - \alpha_t}} \epsilon_\theta(x_t, t)\right) + \sigma_t z_t, \tag{13}$$

where ϵ_θ is a trainable noise predictor, $z_t \sim \mathcal{N}(0, I)$, and $\sigma_t^2 = \beta_t$. The noise predictor is trained by minimizing

$$\mathcal{L}(\theta) = \mathbb{E}_{x,\epsilon_t,t}\left[\|\epsilon_t - \epsilon_\theta(x_t, t)\|_2^2\right], \tag{14}$$

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

Denoising Diffusion Implicit Models and Inversion. Denoising Diffusion Implicit Models (DDIMs) (Song et al., 2021a) modify the forward process into a non-Markovian process. The corresponding denoising update is

$$x_{t-1} = \sqrt{\alpha_{t-1}}\left(\frac{x_t - \sqrt{1-\alpha_t}\,\epsilon_\theta(x_t, t)}{\sqrt{\alpha_t}}\right) + \sqrt{1 - \alpha_{t-1} - \sigma_t^2}\,\epsilon_\theta(x_t, t) + \sigma_t z_t, \tag{15}$$

where $\sigma_t = \eta\sqrt{(1-\alpha_{t-1})/(1-\alpha_t)}\sqrt{1-\alpha_t/\alpha_{t-1}}$. The parameter $\eta \in [0, 1]$ controls stochasticity: $\eta = 1$ recovers DDPM, while $\eta = 0$ yields a deterministic update. The forward process can be modified accordingly, which allows us to deterministically map a clean sample x_0 to a noisy sample x_τ, perform interpolation in the noise space at time τ, and map the result back to the data space.

Naively encoding an image by the stochastic forward process often gives poor reconstructions. DDIM Inversion (Mokady et al., 2023) instead approximately inverts the deterministic denoising process and recovers the noise-space point associated with a given image. Setting $\sigma_t = 0$ in Equation (15) gives

$$x_{t-1} = a_t x_t + b_t \epsilon_\theta(x_t, t) = x_t + (a_t - 1)x_t + b_t\epsilon_\theta(x_t, t), \tag{16}$$

where $a_t = \sqrt{\alpha_{t-1}/\alpha_t}$ and $b_t = -\sqrt{\alpha_{t-1}(1-\alpha_t)/\alpha_t} + \sqrt{1-\alpha_{t-1}}$. This can be regarded as an Euler step for an ordinary differential equation with time derivative $(a_t - 1)x_t + b_t\epsilon_\theta(x_t, t)$. With sufficiently small timestep size, $\epsilon_\theta(x_t, t) \approx \epsilon_\theta(x_{t-1}, t)$, and therefore

$$x_t \approx \frac{x_{t-1} - b_t\epsilon_\theta(x_{t-1}, t)}{a_t} \tag{17}$$

is used as the inverse update. Iterating Equation (17) from $t = 0$ to $t = \tau$ maps a clean sample x_0 to a noisy sample x_τ. Applying the deterministic denoising process from x_τ reconstructs the original x_0 up to numerical error, which substantially improves reconstruction and subsequent interpolation fidelity.

Formulation as Stochastic Differential Equations. As the timestep size approaches zero, the forward process can also be formulated as a stochastic differential equation (SDE) (Song et al., 2021b). The denoising process is the corresponding reverse-time SDE, which depends on the score function $s_\theta(x_t, t) \coloneqq \nabla_{x_t} \log p_t(x_t; \theta)$, where $p_t(x_t; \theta)$ denotes the density of x_t at time t. The noise predictor is closely tied to the score function (Luo, 2022):

$$s_\theta(x_t, t) = \nabla_{x_t} \log p_t(x_t; \theta) \approx -\epsilon_\theta(x_t, t)/\sqrt{1-\alpha_t}. \tag{18}$$

Thus, learning the noise predictor is essentially learning the score function, and the following discussion about s_θ applies to ϵ_θ up to a known scale.

Conditioning and Guidance. The score function can be conditioned on a text prompt c, denoted by $s_\theta(x_t, t, c)$, to guide generation (Rombach et al., 2022). Classifier-free guidance (CFG) amplifies the conditioning to make generated images more faithful to the text prompt (Ho & Salimans, 2021) by replacing the score as

$$\tilde{s}_\theta(x_t, t, c) = (w + 1)s_\theta(x_t, t, c) - w\, s_\theta(x_t, t, \varnothing), \tag{19}$$

where $s_\theta(x_t, t, c)$ and $s_\theta(x_t, t, \varnothing)$ are the conditional and unconditional scores, respectively, and w is the guidance scale. The guidance term can be written as

$$\Delta s(x_t, t, c) = \tilde{s}_\theta(x_t, t, c) - s_\theta(x_t, t, c). \tag{20}$$

Large guidance scales improve prompt alignment but can push samples away from the learned data manifold, motivating the correction in Section 3.3. Negative prompts suppress concepts specified by a complementary prompt c_{neg}:

$$\tilde{s}_\theta(x_t, t, c, c_{\text{neg}}) = s_\theta(x_t, t, c) - w_{\text{neg}} s_\theta(x_t, t, c_{\text{neg}}), \tag{21}$$

where w_{neg} is the negative-prompt scale. When CFG and negative prompts are used simultaneously, the unconditional score in Equation (19) is often replaced by the score conditioned on the negative prompt.

B. Implications of the Proposed Method

B.1. Implication of the Tangent-Minimizing Property

When the score function s_θ is exact, it is the gradient $\nabla_{x_t} \log p_t(x_t;\theta)$ of the log density, and its Jacobian J_{x_t} equals the Hessian $\nabla_{x_t}\nabla_{x_t} \log p_t(x_t;\theta)$. In this idealized case, J_{x_t} is symmetric, and its eigenvectors form an orthonormal basis of the noise space $\mathbb{R}^D$. Following the spectral interpretation of diffusion manifolds (Stanczuk et al., 2024; Ventura et al., 2025), we divide these eigenvectors into a basis for the tangent space $\mathcal{T}_{x_t}\mathcal{M}_t$, $\{v_i\}_{i=1}^d$ with small eigenvalues λ_i, and a basis for the normal space $\mathcal{N}_{x_t}\mathcal{M}_t$, $\{v_j\}_{j=d+1}^D$ with large eigenvalues λ_j. These spaces are orthogonal complements, so the tangent space of the ambient noise space decomposes as $T_{x_t}\mathbb{R}^D = \mathcal{T}_{x_t}\mathcal{M}_t \oplus \mathcal{N}_{x_t}\mathcal{M}_t$. Any tangent vector $v \in T_{x_t}\mathbb{R}^D$ is uniquely decomposed as $v = v_\mathcal{T} + v_\mathcal{N}$, where $v_\mathcal{T} \in \mathcal{T}_{x_t}\mathcal{M}_t$ and $v_\mathcal{N} \in \mathcal{N}_{x_t}\mathcal{M}_t$. The squared Jacobian-vector product can be expanded as

$$\|J_{x_t} v\|_2^2 = \|J_{x_t}(v_\mathcal{T} + v_\mathcal{N})\|_2^2 = \|J_{x_t} v_\mathcal{T}\|_2^2 + \|J_{x_t} v_\mathcal{N}\|_2^2 + 2\langle J_{x_t} v_\mathcal{T}, J_{x_t} v_\mathcal{N}\rangle. \tag{22}$$

Because the eigenspaces are orthogonal, the cross term vanishes, and

$$\|J_{x_t} v_\mathcal{T}\|_2^2 = \sum_{i=1}^{d} \lambda_i^2 \langle v, v_i\rangle^2 \approx 0, \qquad \|J_{x_t} v_\mathcal{N}\|_2^2 = \sum_{j=d+1}^{D} \lambda_j^2 \langle v, v_j\rangle^2 \gg 0 \quad (\text{if } v_\mathcal{N} \neq 0). \tag{23}$$

Thus, minimizing $\|J_{x_t} v\|_2^2$ under a fixed Euclidean norm of v is dominated by minimizing the normal-space component. The proposed metric therefore encourages vectors to lie in the tangent space $\mathcal{T}_{x_t}\mathcal{M}_t$.

In practice, learned score networks are not exactly conservative and J_{x_t} need not be symmetric. Even then, minimizing $\|J_{x_t} v\|_2^2$ suppresses components in the subspace spanned by right singular vectors with large singular values and favors components in the subspace spanned by right singular vectors with small singular values. This is the generalized tangent-normal decomposition used by the proposed metric.

B.2. Regularization and Alternative Construction

The proposed metric $G_{x_t} = J_{x_t}^\top J_{x_t}$ captures the full spectral structure of the score Jacobian, including tangent and normal directions. This is in contrast to the FIM-inspired metric $\lambda s_\theta s_\theta^\top + I$ (Azeglio & Bernardo, 2025). Because the score outer product is rank one, it represents only a single normal direction. The difference is essential when the data manifold has codimension greater than one, which is the typical situation for real data such as images.

One could also consider using J_{x_t} itself as a metric. When the score is conservative, this would correspond to a Hessian manifold induced by the approximate Hessian of the log probability. However, because the log probability is not globally concave, the Hessian can be indefinite. The resulting pseudo-Riemannian structure would no longer characterize geodesics as shortest paths. Using $J_{x_t}^\top J_{x_t}$ avoids this issue while preserving the singular-vector geometry of J_{x_t}.

To guarantee positive definiteness, one can use a regularized metric $G_{x_t} = J_{x_t}^\top J_{x_t} + \lambda I$ for a small $\lambda > 0$. Preliminary experiments with Stable Diffusion v2.1-base (Rombach et al., 2022) showed that this regularization does not materially affect the results, so we use the simpler form in Equation (1).

At $t = 0$, the score function is generally not well trained outside the data manifold, making meaningful path optimization nontrivial. For ideal clean data supported on a low-dimensional manifold, J_{x_t} can also be degenerate at manifold points. In practice, we operate at an intermediate time $\tau > 0$. There, x_τ is perturbed by noise, and J_{x_τ} is empirically full rank with a moderate spectral gap, so the metric behaves as a regular Riemannian metric.

C. Details of Methods

C.1. Comparison Methods

Linear Interpolation. Linear interpolation (LERP) (Ho et al., 2020) treats the noise space at a fixed time $\tau > 0$ as a flat latent space. Given clean endpoints $x_0^{(0)}$ and $x_0^{(1)}$, DDIM Inversion maps them to noisy endpoints $x_\tau^{(0)}$ and $x_\tau^{(1)}$. LERP then constructs

$$x_\tau^{(u)} = (1-u)x_\tau^{(0)} + u x_\tau^{(1)}, \quad u \in [0,1]. \tag{24}$$

Each noisy point $x_\tau^{(u)}$ is mapped back to the data space by the deterministic denoising process.

Algorithm 1 Geodesic-Based Interpolation

1: **Input:** clean endpoints $x_0^{(0)}, x_0^{(1)}$, score s_θ, time τ, points N, iterations K
2: $\{x_\tau^{(0)}, x_\tau^{(1)}\} \leftarrow \{\text{DDIM-Inversion}(x_0^{(i)}; \tau)\}_{i\in\{0,1\}}$
3: Initialize $\{x_\tau^{(u_i)}\}_{i=1}^{N-1}$ by SLERP between $x_\tau^{(0)}$ and $x_\tau^{(1)}$
4: **for** $k = 1$ to K **do**
5: $\quad E \leftarrow \frac{1}{2\Delta u}\sum_{i=0}^{N-1} \|s_\theta(x_\tau^{(u_{i+1})}, \tau) - s_\theta(x_\tau^{(u_i)}, \tau)\|_2^2$
6: $\quad$ Update $\{x_\tau^{(u_i)}\}_{i=1}^{N-1}$ using ∇E with Adam
7: **end for**
8: $\{\hat{x}_0^{(u_i)}\}_{i=0}^{N} \leftarrow \text{DDIM-Denoise}(\{x_\tau^{(u_i)}\}_{i=0}^{N})$
9: **Return:** interpolated clean sequence $\{\hat{x}_0^{(u_i)}\}_{i=0}^{N}$

Algorithm 2 Metric-Based Guidance Correction

1: **Input:** noisy sample x_T, condition c, CFG scale w, weight λ
2: **for** $t = T$ to 1 **do**
3: $\quad$ Compute conditional score $s = s_\theta(x_t, t, c)$
4: $\quad$ Compute guided score $\tilde{s}$ by CFG and set $\Delta s = \tilde{s} - s$
5: $\quad \epsilon \leftarrow \Delta s - (I + \lambda G_{x_t})\Delta s$
6: $\quad \Delta\hat{s}^* \leftarrow \Delta s + \frac{\epsilon^\top \epsilon}{\epsilon^\top (I+\lambda G_{x_t})\epsilon}\epsilon$
7: $\quad$ Update x_{t-1} using $s + \Delta\hat{s}^*$
8: **end for**
9: **Return:** generated sample x_0

Spherical Linear Interpolation. Spherical linear interpolation (SLERP) (Shoemake, 1985; Song et al., 2021a) finds the shortest path on a sphere in the noise space:

$$x_\tau^{(u)} = \frac{\sin((1-u)\theta)}{\sin\theta} x_\tau^{(0)} + \frac{\sin(u\theta)}{\sin\theta} x_\tau^{(1)}, \tag{25}$$

where

$$\theta = \arccos\left(\frac{(x_\tau^{(0)})^\top x_\tau^{(1)}}{\|x_\tau^{(0)}\|_2 \|x_\tau^{(1)}\|_2}\right). \tag{26}$$

SLERP preserves the norms of the noisy samples and often yields more natural interpolations than LERP. However, it assumes that $x_\tau^{(0)}$ and $x_\tau^{(1)}$ are drawn from a normal distribution, which is exact only for sufficiently large τ. It is nevertheless widely used at moderate noise levels and serves as a strong closed-form baseline.

FIM-based Riemannian Metric. Information geometry shows that the Fisher score can induce a Fisher Information Matrix (FIM) on a statistical manifold. Inspired by this, Azeglio & Bernardo (2025) define an FIM-like metric for diffusion models using the score function:

$$g_{x_t}^{\text{FIM}}(v, w) = v^\top(\lambda s_\theta(x_t, t) s_\theta(x_t, t)^\top + I)w. \tag{27}$$

The parameter $\lambda > 0$ balances the FIM term and the Euclidean term. The rank-one term $s_\theta(x_t, t)s_\theta(x_t, t)^\top$ encourages geodesics to be orthogonal to the score direction. However, as discussed in Appendix B.2, it can capture only one normal direction, and therefore geodesics are guaranteed to remain tangent to $\mathcal{M}_t$ only when the data manifold has codimension one. For NAO (Samuel et al., 2023), NoiseDiffusion (Zheng et al., 2024), and GeodesicDiffusion (Yu et al., 2025), we use the default settings from the official implementations.

C.2. Computational Cost of Interpolation Methods

The computational cost varies across interpolation methods. Some methods obtain interpolated images as closed-form solutions, while others require iterative optimization. Table 5 summarizes the costs using the following notation:

- N: number of discretization points for the interpolation path, resulting in $N - 1$ interpolated images.
- S: cost of one score-function, or noise-predictor, evaluation.
- I: number of score evaluations during one DDIM Inversion.
- G: number of score evaluations during one denoising trajectory.
- K: number of optimization iterations for iterative methods.
- L: cost of a simple latent-space operation such as vector addition or scaling, with $L \ll S$.

The total cost consists of three stages: DDIM Inversion of the two endpoints from data space to noise space, interpolation of the endpoints with $N - 1$ intermediate points, and generation by mapping all $N + 1$ noisy points back to data space. LERP, SLERP, and NoiseDiffusion (Zheng et al., 2024) are closed-form methods. NAO (Samuel et al., 2023) is iterative but does not evaluate the score during interpolation, whereas GeodesicDiffusion (Yu et al., 2025), the FIM-based metric (Azeglio & Bernardo, 2025), and ours optimize score-dependent objectives.

Table 5. **Computational cost of interpolation methods.**

Method	Type	Cost
LERP / SLERP / NoiseDiff	closed-form	$2IS + (N-1)L + (N+1)GS$
NAO	iterative	$2IS + K(N+1)L + (N+1)GS$
GeoDiff / FIM-based / Ours	iterative	$2IS + K(N-1)S + (N+1)GS$

C.3. Prompt Adjustment

Following GeodesicDiffusion (Yu et al., 2025), we use prompt adjustment to improve interpolation quality. In Stable Diffusion v2.1-base (Rombach et al., 2022), a text prompt c is first encoded into a CLIP text embedding z (Radford et al., 2021). To better align z with a given image pair $(x_0^{(0)}, x_0^{(1)})$, we adapt the embedding in the same spirit as textual inversion (Gal et al., 2023) by minimizing the DDPM loss in Equation (14). We use AdamW (Loshchilov & Hutter, 2019) with learning rate 0.005 for 500 iterations in image interpolation and 1,000 iterations in video frame interpolation.

Also following Yu et al. (2025), we do not use CFG during interpolation, corresponding to $w = 0$ in Equation (19). We use the negative prompt in Equation (21) with $w_{\text{neg}} = 1$:

> "A doubling image, unrealistic, artifacts, distortions, unnatural blending, ghosting effects, overlapping edges, harsh transitions, motion blur, poor resolution, low detail."

The ablation in Table 6 evaluates the effect of this adjustment.

D. Experimental Setup

This section provides details of the experimental setup in Section 4. Image and video interpolation experiments are conducted on NVIDIA RTX A6000 GPUs. Metric-based guidance correction and distillation experiments are conducted on NVIDIA A100 and H200 GPUs, respectively.

D.1. Synthetic 2D Dataset

Dataset. We construct a two-dimensional C-shaped distribution as follows. We start with an axis-aligned ellipse whose semi-axes are 1.0 along x_1 and 1.2 along x_2. To open the "C", we remove all points in a $\pm 30^\circ$ wedge centered on the positive x_1-axis. We then add isotropic Gaussian perturbations with standard deviation 0.001 per coordinate and draw 100,000 samples from the resulting distribution.

Network. The noise predictor ϵ_θ is a three-layer MLP with hidden width 512 and SiLU activations (Elfwing et al., 2017). The network takes a tuple of a data point x and a normalized time t as input. We set the number of diffusion steps to $T = 1{,}000$ for training and $T = 50$ for generation. The model is trained for 1,000 epochs using AdamW (Loshchilov & Hutter, 2019) with batch size 512. The learning rate follows cosine annealing (Loshchilov & Hutter, 2017), decaying from 10^{-3} to 0 without restarts. For stability, we apply gradient-norm clipping with threshold 1.0.

Implementation Details. For the synthetic visualization, we interpolate between $x_0^{(0)} = (0.0, 1.15)$ and $x_0^{(1)} = (-0.8, -0.6)$ with $N = 100$ discretization points. We use the DDIM scheduler (Song et al., 2021a) and operate in the noise space at $\tau = 0.02T = 1$. For our method and the density-based interpolation baseline of Yu et al. (2025), we find geodesic paths by minimizing the energy functional. Both paths are initialized by SLERP and optimized with Adam (Kingma & Ba, 2015) for 1,000 iterations with learning rate 10^{-4}.

D.2. Datasets for Image Interpolation

For Stable Diffusion v2.1-base (Rombach et al., 2022), we use $T = 50$, $\tau = 0.6T$, and $N - 1 = 9$ interpolated images. DDIM inversion and denoising use no CFG or negative prompt. For geodesic computation, we use the prompt adjustment described in Appendix C.3, and the adjusted score is used to define the metric. For our metric and the FIM-based metric, the path is initialized by SLERP and optimized for 500 Adam iterations with learning rate cosine-annealed from 10^{-3} to 10^{-4}.

MorphBench (Zhang et al., 2023) consists of image pairs obtained via image editing, with 24 pairs in the animation subset

(MB(A)) and 66 pairs in the metamorphosis subset (MB(M)). For MB(A), both endpoints share the same text prompt, which we use as the condition. For MB(M), each endpoint has a distinct prompt; following DiffMorpher (Zhang et al., 2023) and GeodesicDiffusion (Yu et al., 2025), we linearly interpolate the two text embeddings to obtain the condition at each interpolation point.

Animal Faces-HQ (Choi et al., 2020) is a dataset of high-resolution animal-face images. We randomly select 50 dog pairs and 50 cat pairs with LPIPS below 0.6 to ensure semantic similarity. The prompts are "a photo of a dog" and "a photo of a cat," respectively. CelebA-HQ (Karras et al., 2018) is a high-resolution dataset of celebrity faces. We randomly sample 50 male pairs and 50 female pairs, again with LPIPS below 0.6. The prompts are "a photo of a man" and "a photo of a woman," respectively.

D.3. Hyperparameters for Guidance Correction and Distillation

For guidance correction, we use Stable Diffusion v2.1-base with $T = 50$ denoising steps. We generate 30,000 images from MS-COCO 2014 validation captions (Lin et al., 2014) and evaluate FID against the corresponding images and CLIP Score against the text prompts. We set the finite-difference step to $h = 10^{-4}$ and the correction weight to $\lambda = 0.1$.

For distillation, we follow the latent consistency model (LCM) protocol (Luo et al., 2023). The teacher performs a single DDIM step from x_t to x_{t-1}, and the student is trained to minimize the discrepancy between its prediction from x_t and the teacher target obtained from x_{t-1}. The teacher guidance scale is sampled uniformly from $[5, 15]$. The student is parameterized with low-rank adaptation (LoRA) of rank 64 (Hu et al., 2022). We train it for 1,000 iterations on Conceptual 12M (Changpinyo et al., 2021) with batch size 96, Huber loss parameter $c = 0.001$, AdamW learning rate 10^{-4}, and gradient clipping at 1.0. These settings follow the Diffusers defaults (von Platen et al., 2022). After distillation, images are generated with $T = 4$ steps and then evaluated.

E. Additional Results

E.1. Video Frame Interpolation

Experimental Setup. Image-interpolation metrics do not directly compare an interpolated image with a ground-truth middle point. We therefore also evaluate the methods through video frame interpolation, where the middle frame is available and MSE and LPIPS can be computed against ground truth. The goal of this experiment is to objectively evaluate interpolation quality, not to compete with specialized video interpolation methods.

We follow the protocol of Zhu et al. (2024). The benchmark consists of 21 natural-scene clips from DAVIS (Perazzi et al., 2016), 56 human-pose clips from Pexels (Human), and 26 indoor/outdoor clips from RealEstate10K (RE10K) (Zhou et al., 2018). Each clip contains three consecutive frames. Frames 1 and 3 are used as endpoints $x_0^{(0)}$ and $x_0^{(1)}$, while frame 2 is the ground-truth target for $\hat{x}_0^{(0.5)}$. Unless otherwise noted, the model and hyperparameters are the same as in image interpolation. All frames are resized to 512×512, and text prompts are generated from the first frame using BLIP-2 (Li et al., 2022).

Results. Figure 5 shows qualitative video frame interpolation results. The trends match the quantitative results in Table 3: LERP tends to blur, NAO and NoiseDiffusion often deviate substantially from the ground-truth middle frame, and GeoDiff can over-smooth fine texture. In the zoomed Human example, both our method and GeoDiff capture the arm motion, but GeoDiff removes fine structures such as water ripples in DAVIS, indicating that density-based geodesics can sacrifice local detail. Our method preserves edges, object shape, motion, and texture most faithfully across the three datasets.

(continued on p. 15)

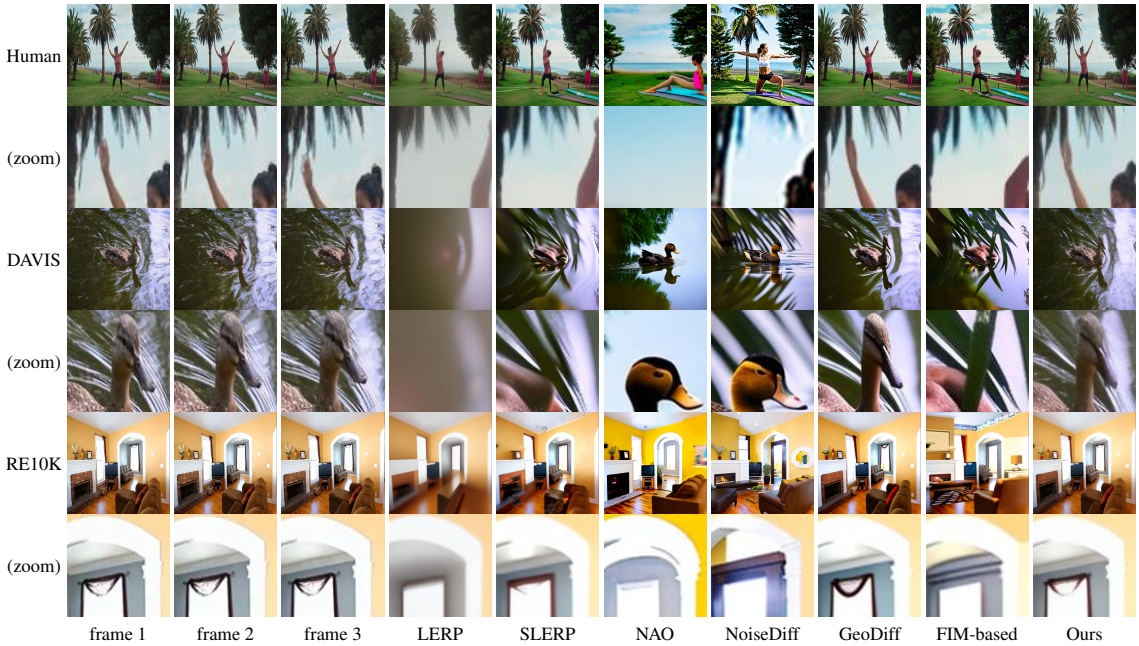

Figure 5. **Qualitative video frame interpolation.** Frame 1 and frame 3 are endpoints, and frame 2 is the ground-truth middle frame.

Table 6. **Ablation of prompt adjustment on video frame interpolation.**

		MSE ↓ ($\times 10^{-3}$)			LPIPS ↓		
Method	Adj.	DAVIS	Human	RE10K	DAVIS	Human	RE10K
SLERP		15.440	6.080	6.128	0.487	0.320	0.301
SLERP	✓	9.894	2.559	3.778	0.355	0.200	0.200
GeoDiff	✓	13.253	3.363	5.941	<u>0.334</u>	<u>0.184</u>	0.229
FIM-based		30.172	11.638	12.679	0.535	0.388	0.373
FIM-based	✓	<u>9.757</u>	<u>2.506</u>	<u>3.001</u>	0.345	0.196	<u>0.194</u>
Ours		13.517	5.008	6.016	0.500	0.350	0.325
Ours	✓	**8.777****	**2.018****	**2.771****	**0.318***	**0.170****	**0.178****

Bold indicates the best result and <u>underline</u> indicates the second best. * 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 ($H_0 : p = 0.5$).

E.2. Ablation Studies

Wall-clock time and cost–quality trade-off. For interpolation, our method has the same asymptotic cost as GeoDiff and the FIM-based metric (Table 5), but takes $1.8\times$ the wall time of GeoDiff on MB(A) because each iteration backpropagates through s_θ at the interpolants to capture local Jacobian structure. This cost is offset by consistent quality gains: our method improves FID on all image datasets and reduces video-interpolation MSE by 34–53% relative to GeoDiff. The iteration ablation in Figure 6 further shows that our method already outperforms GeoDiff and the FIM-based metric at 100 iterations, while the latter two remain below SLERP even at 500 iterations on DAVIS.

Table 7. Wall time

Method	Time (s)
FIM-based	178.29
GeoDiff	216.79
Ours	392.92

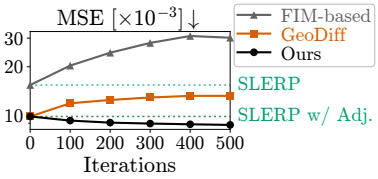

Figure 6. Cost–quality trade-off.

Prompt Adjustment We evaluate the effect of prompt adjustment in video frame interpolation. Because GeoDiff is designed to operate with this adjustment enabled, we omit its no-adjustment variant. Table 6 shows that prompt adjustment improves MSE and LPIPS for SLERP, the FIM-based metric, and our method. With prompt adjustment, our method is best overall, and its gain is larger than that of SLERP. The adjustment helps the guided diffusion model capture a sharper local data manifold, and the proposed metric explicitly leverages this refined local information. By contrast, SLERP is tied to the Gaussian prior geometry and is less sensitive to the refinement.

Noise Level τ and Spectral Gap. The interpolation time τ controls the noisy space in which the path is optimized. Figure 7 visualizes results for varying τ. At $\tau = 0$, a second face appears behind the main face and merges with it, producing an unnatural transition. Without injected noise, the clean data manifold is extremely thin and geodesic optimization under the proposed metric is ill-conditioned. As τ increases, interpolation becomes smoother and more globally coherent. At $\tau = T$, however, the noisy distribution is close to an isotropic Gaussian, the data-manifold structure is washed out, and meaningful geodesics cannot be recovered. Empirically, $\tau \in [0.4T, 0.6T]$ gives the best visual quality.

Figure 8 shows the singular-value distribution of the score Jacobian J_{x_τ}, aggregated over CelebA-HQ images used in our interpolation experiments. Stable Diffusion v2.1-base operates in a VAE latent space of $64 \times 64 \times 4 = 16{,}384$ dimensions. At small τ, hundreds of singular values are near zero, suggesting a local intrinsic dimensionality on the order of a few hundred. The spectral gap between large and small singular values is largest at small τ and decreases as τ grows. At the same time, more singular values approach 1.0 because injected noise thickens the manifold and makes it more isotropic. Moderate τ, such as $0.4T$ or $0.6T$, balances manifold thickness and a clear spectral gap.

Figures 9 and 10 report quantitative scores for different τ. Smaller positive τ often yields better metric values, which does not fully match visual quality; a similar trend was reported by GeodesicDiffusion (Yu et al., 2025). For small τ, interpolation can behave like pixel-wise or patch-wise blending and produce distorted images, while MSE and LPIPS do not necessarily increase substantially. This highlights the need for evaluation metrics that better reflect interpolation quality. In all main experiments, we use $\tau = 0.6T$ for every method except NAO, which is designed to operate at $\tau = T$.

Conjugate Gradient Weight λ and Iterations. We ablate two hyperparameters of the conjugate-gradient solver in Equation (5): the correction weight λ and the number of CG iterations. The evaluation setup follows the guidance-correction experiment in the main text; all other hyperparameters are fixed at their default values, and the CFG scale is set to $w = 7.5$.

Table 8 shows that FID and CLIP Score are nearly constant as λ varies from 10^{-2} to 10^{2}. This robustness follows from the sharp spectral gap of $G_{x_t} = J_{x_t}^\top J_{x_t}$: the one-step CG update suppresses the dominant normal-space eigendirection, whose eigenvalue is much larger than the tangent-space eigenvalues and the identity term.

Table 9 shows the effect of CG iterations. Two iterations further lower FID while nearly preserving CLIP Score, but three iterations degrade performance. Additional iterations refine subdominant eigendirections, but they also accumulate finite-difference and approximate-symmetry errors and require four extra score evaluations per denoising step. We therefore use one CG iteration as the default cost-quality trade-off.

Figure 7. **Qualitative comparison for different interpolation times τ.** Endpoints are shown at both ends of each row.

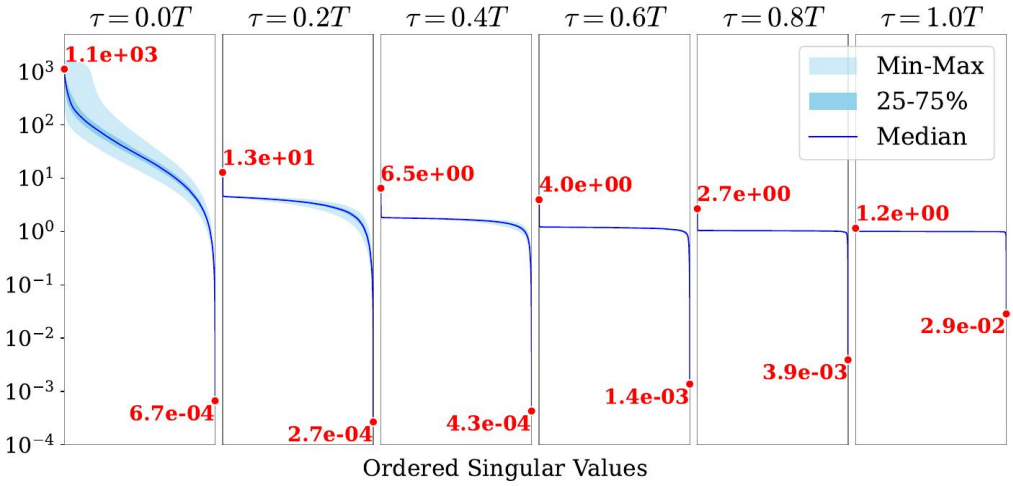

Figure 8. **Singular-value spectrum of the score Jacobian across interpolation times.** The median curve and percentile bands show that smaller τ has a sharper spectral gap, while larger τ makes the noisy distribution more isotropic.

Table 8. **Ablation of the guidance-correction weight λ.**

λ	FID ↓	CLIP ↑
100	13.72	0.314
10	13.80	0.314
1	13.80	0.314
0.1	13.81	0.314
0.01	13.74	0.314

Table 9. **Ablation of the number of CG iterations.**

Iterations	FID ↓	CLIP ↑	NFE ↓
0	14.29	0.314	2
1	13.81	0.314	6
2	13.12	0.313	10
3	23.03	0.308	14

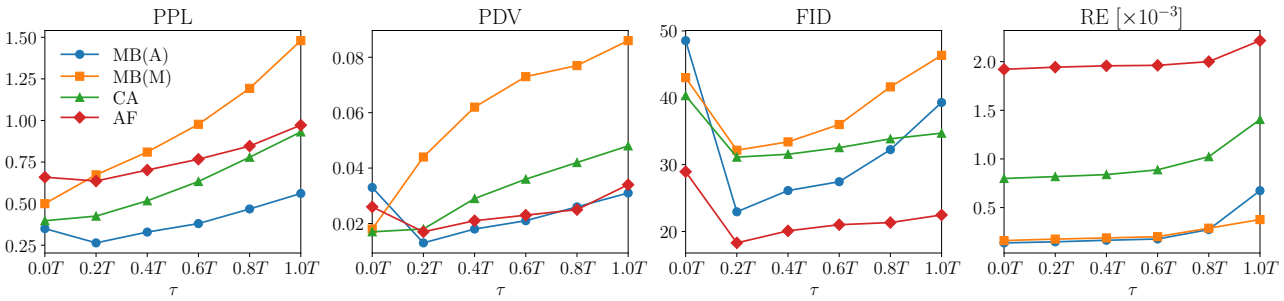

Figure 9. **Ablation of τ on image interpolation metrics.**

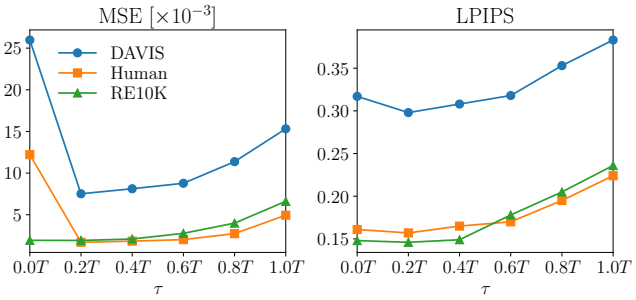

Figure 10. **Ablation of τ on video frame interpolation metrics.**