Zero-shot Geometry-Aware Diffusion Guidance for Music Restoration

Jia-Wei Liao
National Taiwan University
d11922016@csie.ntu.edu.tw

Pin-Chi Pan
National Taiwan University

Li-Xuan Peng
National Tsing Hua University

Sheng-Ping Yang
National Taiwan University

Yen-Tung Yeh
National Taiwan University

Cheng-Fu Chou
National Taiwan University

Yi-Hsuan Yang
National Taiwan University

Abstract

Diffusion models have emerged as powerful generative frameworks and are increasingly used as foundational models for music generation tasks. Recent works have proposed various inference-time optimization methods to adapt pretrained models to downstream tasks. However, these approaches often push noisy samples away from the expected distribution in the diffusion reverse process when applying task-specific loss gradients. To address this issue, we propose Diffusion Geodesic Guidance (DGG), a geometry-aware method that operates on a pretrained diffusion prior preserving the distribution-induced geometry of noisy samples via a closed-form spherical linear interpolation. It updates noisy samples along geodesics of the underlying geometry. We then apply the zero-shot plug-and-play DGG to four multi-task music restoration tasks, achieving consistent improvements over existing training-free baselines and demonstrating a surprisingly wide range of applications for multi-task music restoration.

1 Introduction

Diffusion models [1, 2, 3, 4, 5] have achieved state-of-the-art performance across diverse domains, notably in image [6] and music generation [7, 8, 9, 10, 11, 12]. Their ability to explicitly model complex data distributions makes them promising candidates for downstream music tasks. However, most prior works [13, 14, 15] rely on task-specific supervised training or fine-tuning for each new task, motivating the need for a zero-shot framework that leverages the strong generative priors and generalization capabilities of pretrained diffusion models.

Recent works, spanning music and broader modalities, have explored **inference-time optimization**, which can be broadly divided into two branches. The first branch, exemplified by DITTO [16], focuses on **initial noise optimization**, directly updating the initial noise via backpropagation to align with task-specific objectives. While this can improve objective alignment, it often suffers from vanishing or exploding gradients across the entire sampling trajectory, incurs a high computational cost, and causes the noisy sample to drift away from the standard Gaussian prior, ultimately degrading generation quality. The second branch applies **one-step gradient guidance** at each denoising step. DPS [17] performs gradient descent on noisy samples, which introduces a Jensen gap that can drive samples off the data manifold. In contrast, MPGD [18] assumes that data lie on a linear subspace, allowing updates to preserve the original distribution; however, this assumption is unrealistic in real-world scenarios.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: AI for Music.

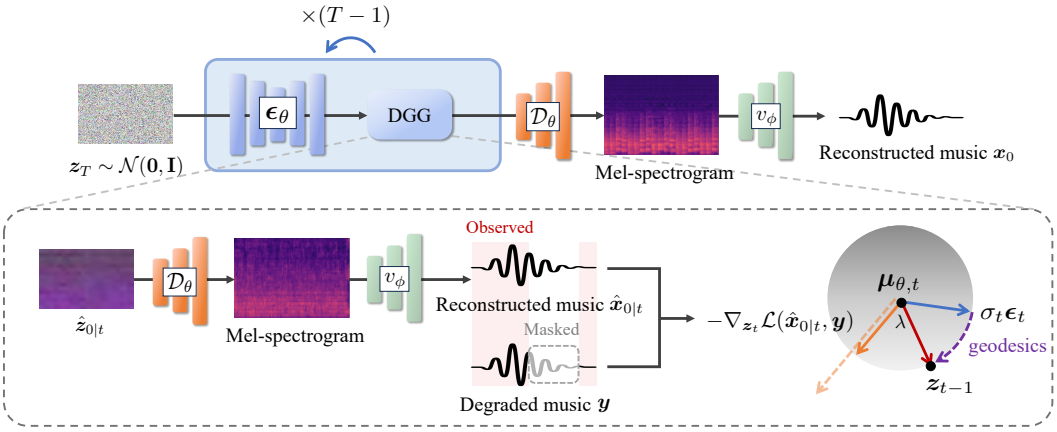

Figure 1: **An overview of the Diffusion Geodesic Guidance (DGG), illustrated using music inpainting.** At each timestep, the intermediate sample $\boldsymbol{z}_t$ is denoised by the latent diffusion model $\boldsymbol{\epsilon}_\theta$ to estimate $\hat{\boldsymbol{z}}_{0|t}$, which is then decoded into a mel-spectrogram by the decoder $\mathcal{D}_\theta$ and further transformed into reconstructed audio $\hat{\boldsymbol{x}}_{0|t}$ via the vocoder v_ϕ. The guidance is computed by evaluating the task-specific loss between $\hat{\boldsymbol{x}}_{0|t}$ and the degraded music $\boldsymbol{y}$, and the intermediate sample $\boldsymbol{z}_{t-1}$ is updated using DGG, which integrates the gradient and sampled noise through spherical interpolation, thereby preventing samples from drifting away from the prior distribution.

To address these issues, we propose Diffusion Geodesic Guidance (DGG), a novel zero-shot guidance method that leverages pretrained diffusion priors at inference time. In DGG, noisy samples are updated along geodesics of the hyperspherical geometry induced by the Gaussian prior. By moving along geodesics via spherical linear interpolation, the updates remain consistent with the underlying geometry, resulting in smoother loss variation during optimization and reducing instability and of pushing samples off the data manifold. DGG integrates seamlessly with any pretrained diffusion model and supports zero-shot music restoration through task-specific losses. Across four music restoration tasks, DGG consistently outperforms recent gradient-based guidance methods, including DPS, MPGD, and DITTO, achieving superior restoration quality with comparable inference speed.

2 Diffusion Geodesic Guidance (DGG)

Overview. We introduce a novel zero-shot diffusion guidance, establishing a unified framework for multi-task music restoration. Each task aims to recover the target waveform from a degraded input $\boldsymbol{y}$. To guide restoration, we define a task-specific loss $\mathcal{L}$ enforcing consistency with the degraded input via an appropriate transformation, introduced in Section 3. We iteratively minimize this loss with pretrained diffusion models at each timestep to reconstruct the restored waveform $\boldsymbol{x}_0$. In the following paragraph, we revisit DDIM and construct a spherical geometry from the reverse diffusion distribution, inspired by the concept of Spherical Gaussians [19], enabling guided updates on the sphere to prevent samples from drifting away from the prior distribution.

DDIM Sampling. Building on Latent Diffusion Models (LDMs) [6, 12, 10], which enable efficient and high-quality generation, we adopt the DDIM formulation [4] with noise schedule $\{\bar{\alpha}_t\}_{t=1}^T$. Sampling begins from $\boldsymbol{z}_T \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$ and proceeds iteratively. At each step, the model $\boldsymbol{\epsilon}_\theta$ estimates the clean sample via Tweedie's formula [20] as $\hat{\boldsymbol{z}}_{0|t} := \frac{1}{\sqrt{\bar{\alpha}_t}}\left(\boldsymbol{z}_t - \sqrt{1-\bar{\alpha}_t}\,\boldsymbol{\epsilon}_\theta(\boldsymbol{z}_t, t)\right)$. The estimate $\hat{\boldsymbol{z}}_{0|t}$ is then used to predict $\boldsymbol{z}_{t-1}$, which can be expressed as $\boldsymbol{z}_{t-1} = \boldsymbol{\mu}_{\theta,t} + \sigma_t\boldsymbol{\epsilon}_t$, where σ_t denotes the noise scale, $\boldsymbol{\epsilon}_t \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$, and the predicted mean $\boldsymbol{\mu}_{\theta,t}$ is given by

$$\boldsymbol{\mu}_{\theta,t} := \sqrt{\bar{\alpha}_{t-1}}\hat{\boldsymbol{z}}_{0|t} + \sqrt{1-\bar{\alpha}_{t-1}-\sigma_t^2}\,\boldsymbol{\epsilon}_\theta(\boldsymbol{z}_t, t).$$

Geometry Induced by the Reverse Diffusion Distribution. To integrate task-specific guidance while preserving consistency with the diffusion process, we induce a spherical geometry from the reverse distribution to regulate the trajectory of latent updates. Given a sampled noise $\boldsymbol{\epsilon}_t \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_n)$, the reverse diffusion distribution $\mathcal{N}(\boldsymbol{\mu}_{\theta,t}, \sigma_t^2\boldsymbol{I}_n)$ induces a sphere centered at $\boldsymbol{\mu}_{\theta,t}$ with

radius proportional to $\sigma_t\|\boldsymbol{\epsilon}_t\|_2$, which can be represented as

$$\mathcal{S}_t := \{\boldsymbol{z} \in \mathbb{R}^n \mid \|\boldsymbol{z} - \boldsymbol{\mu}_{\theta,t}\|_2 = \sigma_t\|\boldsymbol{\epsilon}_t\|_2\}. \tag{1}$$

Diffusion Geodesic Guidance (DGG). Building on the spherical geometry in Equation 1, we design an update rule that ensures each sampling step stays consistent with the reverse diffusion distribution. At each step, $\hat{\boldsymbol{z}}_{0|t}$ is first decoded into a mel-spectrogram by the LDM decoder $\mathcal{D}_\theta$, and then converted into waveform $\hat{\boldsymbol{x}}_{0|t}$ using the vocoder v_ϕ. The task-specific loss $\mathcal{L}(\hat{\boldsymbol{x}}_{0|t}, \boldsymbol{y})$ is subsequently computed, and its gradient $\nabla_{\boldsymbol{z}_t}\mathcal{L}(\hat{\boldsymbol{x}}_{0|t}, \boldsymbol{y})$ serves as the guidance direction. To ensure that the updated sample remains on the sphere $\mathcal{S}_t$, DGG updates $\boldsymbol{z}_{t-1}$ via spherical interpolation between the original noise $\boldsymbol{\epsilon}_t$ and the negative gradient direction $-\nabla_{\boldsymbol{z}_t}\mathcal{L}(\hat{\boldsymbol{x}}_{0|t}, \boldsymbol{y})$ as

$$\boldsymbol{z}_{t-1} = \boldsymbol{\mu}_{\theta,t} + \sigma_t \left(\frac{\sin[(1-\lambda)\beta]}{\sin\beta}\boldsymbol{\epsilon}_t - \frac{\sin(\lambda\beta)}{\sin\beta} \cdot \frac{\|\boldsymbol{\epsilon}_t\|_F}{\|\nabla_{\boldsymbol{z}_t}\mathcal{L}(\hat{\boldsymbol{x}}_{0|t}, \boldsymbol{y})\|_F}\nabla_{\boldsymbol{z}_t}\mathcal{L}(\hat{\boldsymbol{x}}_{0|t}, \boldsymbol{y}) \right),$$

where $\beta = \angle(\boldsymbol{\epsilon}_t, -\nabla_{\boldsymbol{z}_t}\mathcal{L})$ denotes the angle between the noise and gradient directions, and $\lambda \in [0, 1]$ controls the guidance strength. This update steers the latent along the geodesic toward the loss-minimizing direction while remaining constrained to the underlying sphere, thereby preventing deviation from the diffusion reverse distribution. A detailed geometric derivation based on the exponential map is provided in Appendix A. After T steps, we obtain $\boldsymbol{z}_0$, which is decoded into the final waveform $\boldsymbol{x}_0$ through $\mathcal{D}_\theta$ and v_ϕ. The overall framework is illustrated in Figure 1, and the complete sampling procedure is detailed in Algorithm 4.

3 Applying to Music Restoration Tasks

We evaluate our method on four music restoration tasks, each formulated with a corresponding degradation process and task specific loss $\mathcal{L}$ with the log-mel spectrogram transform ψ.

Inpainting. A binary mask $\boldsymbol{M}$ marks the missing region of the clean waveform $\boldsymbol{x}$gt over $[t_{\text{start}}, t_{\text{end}}]$, producing the degraded signal $\boldsymbol{y} = \boldsymbol{M} \odot \boldsymbol{x}_{\text{gt}}$. The goal is to reconstruct $\boldsymbol{x}_{\text{gt}}$ by minimizing:

$$\mathcal{L}_{\text{Inpaint}}(\boldsymbol{x}, \boldsymbol{y}) = \|\psi(\boldsymbol{M} \odot \boldsymbol{x}) - \psi(\boldsymbol{y})\|_F.$$

Super-Resolution. A low-resolution music $\boldsymbol{y}$ is generated from $\boldsymbol{x}_{\text{gt}}$ via a downsampling operator $\boldsymbol{D}_r$ with rate r, i.e., $\boldsymbol{y} = \boldsymbol{D}_r(\boldsymbol{x}_{\text{gt}})$. The goal is to recover the high-resolution $\boldsymbol{x}_{\text{gt}}$ by minimizing:

$$\mathcal{L}_{\text{SR}}(\boldsymbol{x}, \boldsymbol{y}) = \|\psi(\boldsymbol{D}_r(\boldsymbol{x})) - \psi(\boldsymbol{y})\|_F.$$

Dereverberation. Reverberation is simulated by convolving $\boldsymbol{x}_{\text{gt}}$ with a Room Impulse Response (RIR) $\boldsymbol{h}$, yielding $\boldsymbol{y} = \boldsymbol{h} * \boldsymbol{x}_{\text{gt}}$. The objective is to recover the dry music $\boldsymbol{x}_{\text{gt}}$ by minimizing:

$$\mathcal{L}_{\text{Derev}}(\boldsymbol{x}, \boldsymbol{y}) = \|\psi(\boldsymbol{h} * \boldsymbol{x}) - \psi(\boldsymbol{y})\|_F.$$

Phase Retrieval. The magnitude spectrogram $\boldsymbol{y} = |\boldsymbol{F}(\boldsymbol{x}\text{gt})|$ is obtained via the Short-Time Fourier Transform (STFT) $\boldsymbol{F}$, discarding the phase. The goal is to reconstruct $\boldsymbol{x}_{\text{gt}}$ from $\boldsymbol{y}$ by minimizing:

$$\mathcal{L}_{\text{PR}}(\boldsymbol{x}, \boldsymbol{y}) = \|\psi(|\boldsymbol{F}(\boldsymbol{x})|) - \psi(\boldsymbol{y})\|_F.$$

4 Experiments

4.1 Implementation Details

We sample 100 music tracks from the MoisesDB [21] and MusicCaps [22] datasets to construct evaluation subsets. MoisesDB provides isolated instrument stems and structured mixtures, enabling fine-grained analysis of instrument-specific restoration performance. In contrast, MusicCaps offers broader stylistic diversity and rich textual annotations, making it well-suited for assessing model generalization and exploring prompt-based conditioning. From each dataset, we randomly extract non-overlapping 5-second segments from the original tracks, which are used consistently across all restoration tasks. We adopt AudioLDM2 [12] as the pretrained diffusion backbone and process all audio at a sampling rate of 16 kHz. Log-mel spectrograms are computed using a 1024-point FFT, a hop size of 160 samples, and 64 mel frequency bins, yielding a time–frequency representation compatible with the model's input format. The denoising process follows the DDIM sampler with 500 steps, keeping all model parameters frozen during inference. A null-text prompt is used as the conditioning input.

Dataset	Inpainting		Super-Resolution		Dereverberation		Phase Retrieval	
	LSD ↓	FAD ↓	LSD ↓	FAD ↓	LSD ↓	FAD ↓	LSD ↓	FAD ↓
MoisesDB								
DPS [17]	<u>0.7960</u>	<u>0.4847</u>	<u>1.1019</u>	<u>0.5794</u>	<u>1.1985</u>	0.6482	**0.6973**	<u>0.5325</u>
MPGD [18]	1.7190	0.6108	1.7423	0.6096	1.7386	<u>0.6011</u>	1.7197	0.6018
DITTO [16]	1.1250	0.7284	1.1610	0.8863	1.2164	0.9149	0.8551	0.8353
DGG (Ours)	**0.6363**	**0.2904**	**0.8897**	**0.4341**	**1.0582**	**0.4444**	<u>0.7056</u>	**0.4666**
MusicCaps								
DPS [17]	<u>0.9026</u>	0.5185	<u>1.0453</u>	<u>0.4727</u>	<u>1.0449</u>	<u>0.5000</u>	**0.7388**	0.5683
MPGD [18]	1.2734	<u>0.5153</u>	1.2864	0.5151	1.2815	0.5115	1.2632	<u>0.5142</u>
DITTO [16]	1.2304	0.8514	1.5580	0.9673	1.5316	0.7463	1.2598	0.9125
DGG (Ours)	**0.7019**	**0.2597**	**0.9617**	**0.3244**	**0.9322**	**0.3278**	<u>0.8051</u>	**0.4430**

Table 1: Quantitative results on music restoration tasks for two datasets. LSD and FAD are reported separately. The best is marked in **bold** and the second best is <u>underlined</u>.

4.2 Results

In Table 1, we report quantitative results for four music restoration tasks on MoisesDB and MusicCaps datasets using AudioLDM2 [12]. Our method, DGG, consistently outperforms DPS [17], MPGD [18], and DITTO [16] in both spectral distortion (LSD) and perceptual quality (FAD). Compared to DITTO, DGG achieves over 5–6× faster inference time while maintaining comparable efficiency to DPS and MPGD. In addition, DITTO performs full-trajectory initial-noise optimization, which suffers from instability and exploding gradients, often degrading output quality despite its high computational cost. These results show that DGG not only restores music with high fidelity, as indicated by lower LSD, but also enhances perceptual realism, as reflected in improved FAD, which is crucial for real-world listening scenarios. In qualitative comparisons, DGG produces reconstructions that are most consistent with the ground truth in both spectral structure and spectral energy. In contrast, MPGD [18] fails to reconstruct coherent harmonic patterns, resulting in overly smeared and noisy outputs. DITTO [16] recovers low-frequency components but lacks detail in the mid-to-high frequency range. DPS [17] performs reasonably well but introduces artifacts in the reconstructed region.

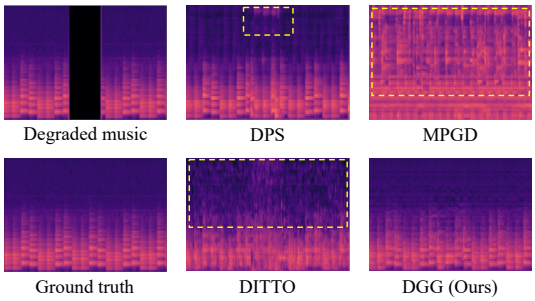

Figure 2: Qualitative comparison of mel-spectrograms for the inpainting task. Dashed boxes highlight regions with noticeable generation or reconstruction artifacts.

5 Conclusion

We propose Diffusion Geodesic Guidance (DGG), a geometry-aware, zero-shot guidance framework that updates noisy samples along geodesics of the hyperspherical geometry induced by the Gaussian prior. By leveraging spherical linear interpolation, DGG preserves the distribution-induced geometry throughout the denoising process, mitigating instability and preventing samples from drifting off the data manifold. Experiments on four music restoration tasks show that DGG consistently outperforms recent inference-time optimization methods, achieving state-of-the-art performance in both LSD and FAD with comparable inference speed.

Acknowledgements

This research is supported by National Science and Technology Council, Taiwan (R.O.C) under the grant numbers 114-2221-E-002-182-MY3 and 113-2221-E-002-201.

References

[1] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning (ICML)*, 2015.

[2] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[3] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations (ICLR)*, 2020.

[4] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations (ICLR)*, 2021.

[5] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[6] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[7] Seth Forsgren and Hayk Martiros. Riffusion-stable diffusion for real-time music generation, 2022.

[8] Flavio Schneider, Ojasv Kamal, Zhijing Jin, and Bernhard Schölkopf. Moûsai: Text-to-music generation with long-context latent diffusion. *arXiv preprint*, 2023.

[9] Haohe Liu, Zehua Chen, Yi Yuan, Xinhao Mei, Xubo Liu, Danilo Mandic, Wenwu Wang, and Mark D Plumbley. AudioLDM: Text-to-audio generation with latent diffusion models. In *International Conference on Machine Learning (ICML)*, 2023.

[10] Ke Chen, Yusong Wu, Haohe Liu, Marianna Nezhurina, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. MusicLDM: Enhancing novelty in text-to-music generation using beat-synchronous mixup strategies. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024.

[11] Zach Evans, CJ Carr, Josiah Taylor, Scott H Hawley, and Jordi Pons. Fast timing-conditioned latent audio diffusion. In *International Conference on Machine Learning (ICML)*, 2024.

[12] Haohe Liu, Yi Yuan, Xubo Liu, Xinhao Mei, Qiuqiang Kong, Qiao Tian, Yuping Wang, Wenwu Wang, Yuxuan Wang, and Mark D Plumbley. AudioLDM 2: Learning holistic audio generation with self-supervised pretraining. In *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLPRO)*, 2024.

[13] Lejun Min, Junyan Jiang, Gus Xia, and Jingwei Zhao. Polyffusion: A diffusion model for polyphonic score generation with internal and external controls. In *International Society for Music Information Retrieval (ISMIR)*, 2023.

[14] Eloi Moliner, Jaakko Lehtinen, and Vesa Välimäki. Solving audio inverse problems with a diffusion model. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023.

[15] Haohe Liu, Ke Chen, Qiao Tian, Wenwu Wang, and Mark D Plumbley. Audiosr: Versatile audio super-resolution at scale. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024.

[16] Zachary Novack, Julian McAuley, Taylor Berg-Kirkpatrick, and Nicholas J Bryan. DITTO: Diffusion inference-time t-optimization for music generation. In *International Conference on Machine Learning (ICML)*, 2024.

[17] Hyungjin Chung, Jeongsol Kim, Michael Thompson Mccann, Marc Louis Klasky, and Jong Chul Ye. Diffusion posterior sampling for general noisy inverse problems. In *International Conference on Learning Representations (ICLR)*, 2023.

[18] Yutong He, Naoki Murata, Chieh-Hsin Lai, Yuhta Takida, Toshimitsu Uesaka, Dongjun Kim, Wei-Hsiang Liao, Yuki Mitsufuji, J Zico Kolter, Ruslan Salakhutdinov, et al. Manifold preserving guided diffusion. In *International Conference on Learning Representations (ICLR)*, 2024.

[19] Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*, volume 47. Cambridge University Press, 2018.

[20] Bradley Efron. Tweedie's formula and selection bias. *Journal of the American Statistical Association*, 2011.

[21] Igor G Pereira, Felipe Araujo, Filip Korzeniowski, and Richard Vogl. Moisesdb: A dataset for source separation beyond 4 stems. In *International Society for Music Information Retrieval (ISMIR)*, 2023.

[22] Andrea Agostinelli, Timo I Denk, Zalán Borsos, Jesse Engel, Mauro Verzetti, Antoine Caillon, Qingqing Huang, Aren Jansen, Adam Roberts, Marco Tagliasacchi, et al. Musiclm: Generating music from text. *arXiv preprint arXiv:2301.11325*, 2023.

[23] Augustine Gray and John Markel. Distance measures for speech processing. In *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLPRO)*, 1976.

[24] Dominik Roblek, Kevin Kilgour, Matt Sharifi, and Mauricio Zuluaga. Fréchet audio distance: A reference-free metric for evaluating music enhancement algorithms. In *Conference of the International Speech Communication Association (INTERSPEECH)*, 2019.

[25] Azalea Gui, Hannes Gamper, Sebastian Braun, and Dimitra Emmanouilidou. Adapting frechet audio distance for generative music evaluation. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024.

Appendix

A Derivation of Diffusion Geodesic Guidance (DGG)

In this section, we derive the DGG update by interpreting guidance as a Riemannian gradient step constrained to the diffusion sphere. This geometric viewpoint leads naturally to an exponential map formulation, yielding a closed-form spherical interpolation between the noise and task gradient directions.

Riemannian Gradient Direction. To remain on $\mathcal{S}_t$, as defined in Equation 1, we move along the tangent space at $\boldsymbol{z}_t$. The orthogonal projection of $\nabla_{\boldsymbol{z}_t}\mathcal{L}$ onto tangent space is

$$\operatorname{grad}\mathcal{L} := \nabla_{\boldsymbol{z}_t}\mathcal{L} - \frac{\langle \nabla_{\boldsymbol{z}_t}\mathcal{L}, \boldsymbol{\epsilon}_t\rangle}{\|\boldsymbol{\epsilon}_t\|_2^2}\boldsymbol{\epsilon}_t.$$

Let $\widehat{\operatorname{grad}\mathcal{L}} := \operatorname{grad}\mathcal{L}/\|\operatorname{grad}\mathcal{L}\|_2$ be the unit tangent descent direction. Define the angle $\beta := \angle(\boldsymbol{\epsilon}_t, -\nabla_{\boldsymbol{z}_t}\mathcal{L}) \in [0,\pi]$. Then

$$\widehat{\operatorname{grad}\mathcal{L}} = \frac{1}{\sin\beta}\left(\frac{\nabla_{\boldsymbol{z}_t}\mathcal{L}}{\|\nabla_{\boldsymbol{z}_t}\mathcal{L}\|_2} + \cos\beta\frac{\boldsymbol{\epsilon}_t}{\|\boldsymbol{\epsilon}_t\|_2}\right).$$

Exponential Map Update on the Sphere. A geodesic step of size $\eta > 0$ along $-\widehat{\operatorname{grad}\mathcal{L}}$ on the sphere of radius $\sigma_t\|\boldsymbol{\epsilon}_t\|_2$ centered at $\boldsymbol{\mu}_{\theta,t}$ is

$$\begin{aligned}
\boldsymbol{z}_{t-1} &= \operatorname{Exp}_{\boldsymbol{z}_t}(-\eta\widehat{\operatorname{grad}\mathcal{L}}) \\
&= \boldsymbol{\mu}_{\theta,t} + \cos\left(\frac{\eta}{\sigma_t\|\boldsymbol{\epsilon}_t\|_2}\right)\sigma_t\boldsymbol{\epsilon}_t - \sin\left(\frac{\eta}{\sigma_t\|\boldsymbol{\epsilon}_t\|_2}\right)\sigma_t\|\boldsymbol{\epsilon}_t\|_2\widehat{\operatorname{grad}\mathcal{L}} \\
&= \boldsymbol{\mu}_{\theta,t} + \sigma_t\left[\left(\cos\left(\frac{\eta}{\sigma_t\|\boldsymbol{\epsilon}_t\|_2}\right) - \cot\beta\sin\left(\frac{\eta}{\sigma_t\|\boldsymbol{\epsilon}_t\|_2}\right)\right)\boldsymbol{\epsilon}_t - \frac{\sin[\eta/(\sigma_t\|\boldsymbol{\epsilon}_t\|_2)]}{\sin\beta}\cdot\frac{\|\boldsymbol{\epsilon}_t\|_2}{\|\nabla_{\boldsymbol{z}_t}\mathcal{L}\|_2}\nabla_{\boldsymbol{z}_t}\mathcal{L}\right].
\end{aligned}$$

From Exponential Map to Spherical Linear Interpolation. Choose the step size to match the geodesic interpolation parameter: set

$$\frac{\eta}{\sigma_t\|\boldsymbol{\epsilon}_t\|_2} \;=\; \lambda\beta \text{ with } \lambda \in [0,1].$$

Then

$$\cos\left(\frac{\eta}{\sigma_t\|\boldsymbol{\epsilon}_t\|_2}\right) - \cot\beta\sin\left(\frac{\eta}{\sigma_t\|\boldsymbol{\epsilon}_t\|_2}\right) = \frac{\sin[(1-\lambda)\beta]}{\sin\beta} \text{ and } \frac{\sin[\eta/(\sigma_t\|\boldsymbol{\epsilon}_t\|_2)]}{\sin\beta} = \frac{\sin(\lambda\beta)}{\sin\beta}.$$

Therefore, we have

$$\boldsymbol{z}_{t-1} = \boldsymbol{\mu}_{\theta,t} + \sigma_t\left(\frac{\sin[(1-\lambda)\beta]}{\sin\beta}\boldsymbol{\epsilon}_t - \frac{\sin(\lambda\beta)}{\sin\beta}\cdot\frac{\|\boldsymbol{\epsilon}_t\|_2}{\|\nabla_{\boldsymbol{z}_t}\mathcal{L}\|_2}\nabla_{\boldsymbol{z}_t}\mathcal{L}\right).$$

This is exactly spherical linear interpolation on $\mathcal{S}_t$ between the current direction $\boldsymbol{\epsilon}_t/\|\boldsymbol{\epsilon}_t\|_2$ and the gradient direction $-\nabla_{\boldsymbol{z}_t}\mathcal{L}/\|\nabla_{\boldsymbol{z}_t}\mathcal{L}\|_2$, with interpolation parameter $\lambda \in [0,1]$.

B More Implementation Details

B.1 Algorithms of Baseline Methods

We adopt diffusion-based music foundation models, using AudioLDM2 [12] as the backbone, and compare our method against several zero-shot baselines: DITTO [16], DPS [17] and MPGD [18]). For a fair comparison, we use the best-performing hyperparameters reported for each baseline: a

guidance strength of 5×10^{-4} for DPS, 5×10^{-3} for MPGD, and a learning rate of 0.5 for DITTO. For our proposed DGG, we set the guidance strength to 0.08, which we found to be the best-performing parameter based on our experiments. For DITTO, we follow its original setup with 20 diffusion timesteps, gradient descent with 100 inner loop. The detailed algorithms for DPS are provided in Algorithm 1, for MPGD in Algorithm 2, for DITTO in Algorithm 3 and for DGG in Algorithm 4.

Algorithm 1 DPS [17]

Input: Degraded music $\boldsymbol{y}$, UNet $\boldsymbol{\epsilon}_\theta$, VAE decoder $\mathcal{D}_\theta$, Vocoder v_ϕ, DDIM parameters $\bar{\alpha}_t$, Noise levels σ_t, Task specific loss $\mathcal{L}$, Guidance strength $\gamma > 0$.
$\boldsymbol{z}_T \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$.
for $t = T$ to 1 **do**
  $\boldsymbol{\epsilon}_t \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$.
  $\hat{\boldsymbol{z}}_{0|t} \leftarrow \frac{1}{\sqrt{\bar{\alpha}_t}} \left(\boldsymbol{z}_t - \sqrt{1-\bar{\alpha}_t} \boldsymbol{\epsilon}_\theta(\boldsymbol{z}_t, t) \right)$.
  $\hat{\boldsymbol{z}}_{t-1} \leftarrow \sqrt{\bar{\alpha}_{t-1}} \hat{\boldsymbol{z}}_{0|t} + \sqrt{1-\bar{\alpha}_{t-1}-\sigma_t^2} \boldsymbol{\epsilon}_\theta(\boldsymbol{z}_t, t) + \sigma_t \boldsymbol{\epsilon}_t$.
  $\hat{\boldsymbol{x}}_{0|t} \leftarrow (v_\phi \circ \mathcal{D}_\theta)(\hat{\boldsymbol{z}}_{0|t})$.
  $\boldsymbol{z}_{t-1} \leftarrow \hat{\boldsymbol{z}}_{t-1} - \gamma \nabla_{\boldsymbol{z}_t} \mathcal{L}(\hat{\boldsymbol{x}}_{0|t}, \boldsymbol{y})$.
end for
return $(v_\phi \circ \mathcal{D}_\theta)(\boldsymbol{z}_0)$.

Algorithm 2 MPGD [18]

Input: Degraded music $\boldsymbol{y}$, UNet $\boldsymbol{\epsilon}_\theta$, VAE decoder $\mathcal{D}_\theta$, Vocoder v_ϕ, DDIM parameters $\bar{\alpha}_t$, Noise levels σ_t, Task specific loss $\mathcal{L}$, Guidance strength $\gamma > 0$.
$\boldsymbol{z}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.
for $t = T$ to 1 **do**
  $\boldsymbol{\epsilon}_t \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$.
  $\hat{\boldsymbol{z}}_{0|t} \leftarrow \frac{1}{\sqrt{\bar{\alpha}_t}} \left(\boldsymbol{z}_t - \sqrt{1-\bar{\alpha}_t} \boldsymbol{\epsilon}_\theta(\boldsymbol{z}_t, t) \right)$.
  $\hat{\boldsymbol{x}}_{0|t} \leftarrow (v_\phi \circ \mathcal{D}_\theta)(\hat{\boldsymbol{z}}_{0|t})$.
  $\hat{\boldsymbol{z}}^*_{0|t} \leftarrow \hat{\boldsymbol{z}}_{0|t} - \gamma \nabla_{\hat{\boldsymbol{z}}_{0|t}} \mathcal{L}(\hat{\boldsymbol{x}}_{0|t}, \boldsymbol{y})$.
  $\boldsymbol{z}_{t-1} \leftarrow \sqrt{\bar{\alpha}_{t-1}} \hat{\boldsymbol{z}}^*_{0|t} + \sqrt{1-\bar{\alpha}_{t-1}-\sigma_t^2} \left(\frac{\boldsymbol{z}_t - \sqrt{\bar{\alpha}_t} \hat{\boldsymbol{z}}^*_{0|t}}{\sqrt{1-\bar{\alpha}_t}} \right) + \sigma_t \boldsymbol{\epsilon}_t$.
end for
return $(v_\phi \circ \mathcal{D}_\theta)(\boldsymbol{z}_0)$.

Algorithm 3 DITTO [16]

Input: Degraded music $\boldsymbol{y}$, UNet $\boldsymbol{\epsilon}_\theta$, VAE decoder $\mathcal{D}_\theta$, Vocoder v_ϕ, DDIM parameters $\bar{\alpha}_t$, Noise levels σ_t, Task specific loss $\mathcal{L}$, Guidance strength $\gamma > 0$.
$\boldsymbol{z}_T^{(0)} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$.
for $k = 0$ to $K-1$ **do**
  for $t = T$ to 1 **do**
    $\boldsymbol{\epsilon}_t^{(k)} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$.
    $\boldsymbol{z}_{t-1}^{(k)} \leftarrow \frac{\sqrt{\bar{\alpha}_{t-1}}}{\sqrt{\bar{\alpha}_t}} \left(\boldsymbol{z}_t^{(k)} - \sqrt{1-\bar{\alpha}_t} \boldsymbol{\epsilon}_\theta(\boldsymbol{z}_t^{(k)}, t) \right) + \sqrt{1-\bar{\alpha}_{t-1}-\sigma_t^2} \boldsymbol{\epsilon}_\theta(\boldsymbol{z}_t^{(k)}, t) + \sigma_t \boldsymbol{\epsilon}_t^{(k)}$.
  end for
  $\boldsymbol{x}_0^{(k)} \leftarrow (v_\phi \circ \mathcal{D}_\theta)(\boldsymbol{z}_0^{(k)})$.
  $\boldsymbol{z}_T^{(k+1)} \leftarrow \boldsymbol{z}_T^{(k)} - \gamma \nabla_{\boldsymbol{z}_T^{(k)}} \mathcal{L}(\hat{\boldsymbol{x}}_0^{(k)}, \boldsymbol{y})$.
end for
return $\boldsymbol{x}_0^{(K-1)}$.

Algorithm 4 DGG (Ours)

Input: Degraded music $\boldsymbol{y}$, UNet $\boldsymbol{\epsilon}_\theta$, VAE decoder $\mathcal{D}_\theta$, Vocoder v_ϕ, DDIM parameters $\bar{\alpha}_t$, Noise levels σ_t, Task specific loss $\mathcal{L}$, Guidance strength $0 \leq \lambda \leq 1$.
$\boldsymbol{z}_T \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$.
for $t = T$ to 1 **do**
  $\hat{\boldsymbol{z}}_{0|t} \leftarrow \frac{1}{\sqrt{\bar{\alpha}_t}} \left(\boldsymbol{z}_t - \sqrt{1-\bar{\alpha}_t} \boldsymbol{\epsilon}_\theta(\boldsymbol{z}_t, t) \right)$.
  $\boldsymbol{\mu}_{\theta,t} \leftarrow \sqrt{\bar{\alpha}_{t-1}} \hat{\boldsymbol{z}}_{0|t} + \sqrt{1-\bar{\alpha}_{t-1}-\sigma_t^2}\, \boldsymbol{\epsilon}_\theta(\boldsymbol{z}_t, t)$.
  $\hat{\boldsymbol{x}}_{0|t} \leftarrow (v_\phi \circ \mathcal{D}_\theta)(\hat{\boldsymbol{z}}_{0|t})$.
  $\boldsymbol{g} \leftarrow \nabla_{\boldsymbol{z}_t} \mathcal{L}(\hat{\boldsymbol{x}}_{0|t}, \boldsymbol{y})$.
  $\boldsymbol{\epsilon}_t \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$.
  $\beta \leftarrow \angle(\boldsymbol{\epsilon}_t, -\boldsymbol{g})$.
  $\boldsymbol{z}_{t-1} \leftarrow \boldsymbol{\mu}_{\theta,t} + \sigma_t \left(\frac{\sin[(1-\lambda)\beta]}{\sin\beta} \boldsymbol{\epsilon}_t - \frac{\sin(\lambda\beta)}{\sin\beta} \cdot \frac{\|\boldsymbol{\epsilon}_t\|_F}{\|\boldsymbol{g}\|_F} \boldsymbol{g} \right)$.
end for
return $(v_\phi \circ \mathcal{D}_\theta)(\boldsymbol{z}_0)$.

B.2 Experimental Setup for Music Restoration

We evaluate DGG on four music restoration tasks with the following configurations. (i) **Music Inpainting** masks the segment between 2–3 seconds without downsampling (scale = 1), with three masking strategies: fixed box (2–3s), random (mask ratio 0.3), and periodic (interval 1s, mask duration 0.1s). (ii) **Super-Resolution** downsamples the audio by a factor of 2 before reconstruction. (iii) **Phase Retrieval** reconstructs the phase from the magnitude spectrogram, using the same spectrogram configuration as in the global setup, namely a 1024-point FFT with a hop size of 160 and a window length of 1024. (iv) **Music Dereverberation** applies simulated reverberation with an impulse response length of 5000 and a decay factor of 0.99. Here, the reverberation is synthetically added and does not correspond to natural recording or mixing conditions, so its removal can be viewed as eliminating an artificial degradation. All experiments are implemented in PyTorch and executed on a single NVIDIA GeForce RTX 4090 GPU.

B.3 Evaluation Metrics

We employ two complementary metrics to evaluate restoration quality: Log-Spectral Distance (LSD) [23] in the frequency domain, and Fréchet Audio Distance (FAD) [24, 25] with the CLAP music backbone to measure perceptual similarity at the distribution level. Both metrics are applied uniformly across all restoration tasks to ensure consistent and fair comparison.

C Broader Impacts and Limitation

Music restoration is crucial for enhancing archival recordings, improving user-generated content, and enabling interactive music editing tools. Our proposed DGG method is a plug-and-play approach that can be applied to any pretrained music diffusion model for music restoration. It is computationally efficient and avoids the expensive cost of supervised training, making restoration methods more accessible for both research and creative applications. However, its effectiveness is inherently bounded by the capability of the underlying pretrained music diffusion model, meaning that if the backbone model lacks sufficient representation power for certain genres or instruments, the restoration performance may degrade accordingly.