39th Conference on Neural Information Processing System AI for Music Workshop

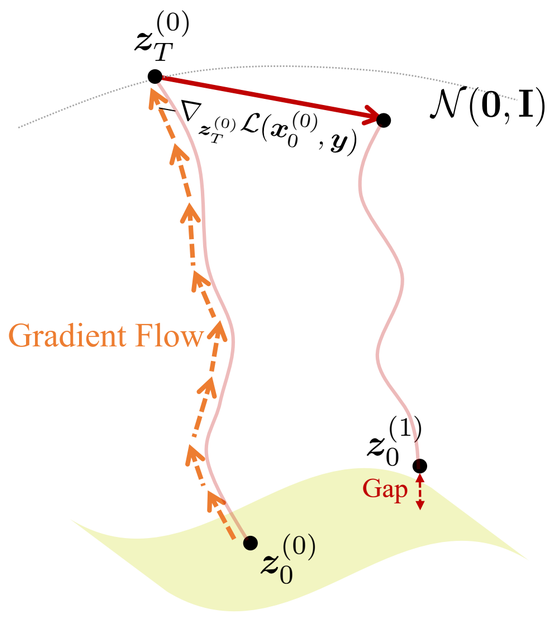
Zero-shot Geometry-Aware Diffusion Guidance for Music Restoration

Jia-Wei Liao[1], Pin-Chi Pan[1], Li-Xuan Peng[2], Sheng-Ping Yang[1], Yen-Tung Yeh[1],
Cheng-Fu Chou[1], Yi-Hsuan Yang[1]
[1] National Taiwan University, [2] National Tsing Hua University

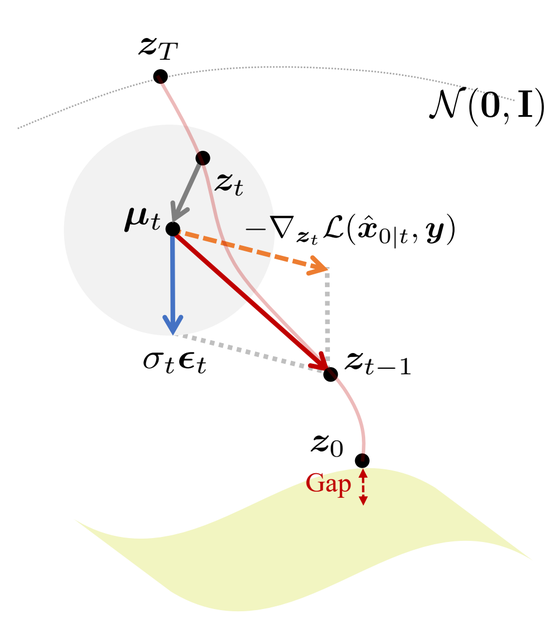
Restore music through the lens of diffusion geometry, no training required. DGG is a geometry-aware plug-in that works with any pretrained diffusion model.

Motivation

- Diffusion models provide strong generative priors and work well for music.
- When applying pretrained diffusion models to music restoration tasks, existing approaches often require task-specific supervised training or fine-tuning for each new task, which limits generalization and scalability.
- Prior inference-time optimization avoids retraining but still has key issues:
  a) **Initial Noise Optimization:** suffers from vanishing/exploding gradients, high computation, and drift from the Gaussian prior.
  b) **Gradient-based Stepwise Guidance:** often pushes noisy sample off the data manifold because it doesn't respect the latent geometry.

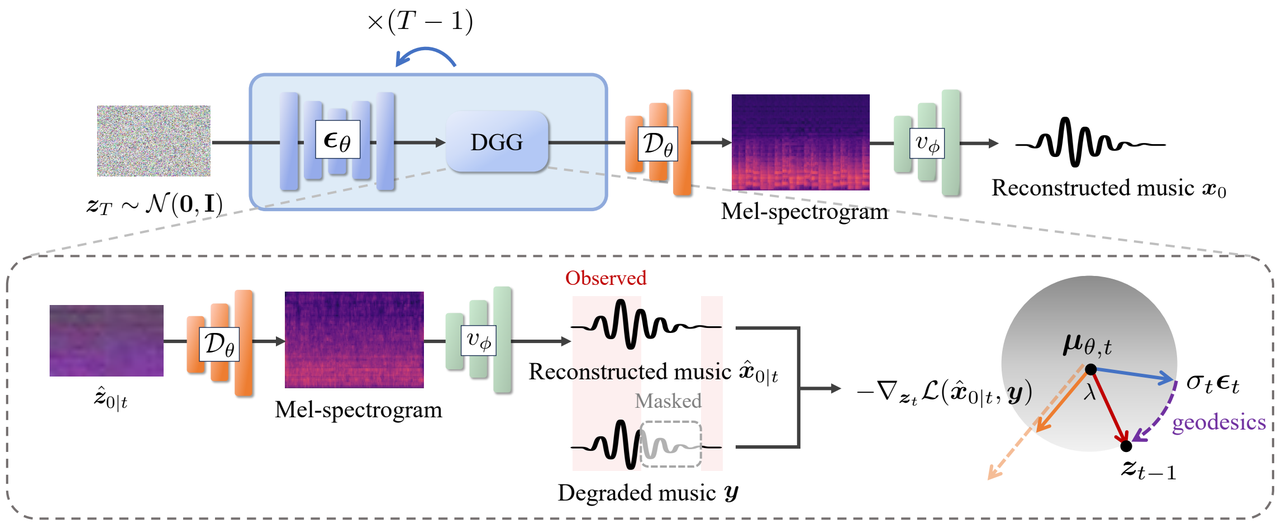


(a) Initial Noise Optimization



(b) Gradient-based Stepwise Guidance

Diffusion Geodesic Guidance (DGG)



Geometry Induced by the Reverse Diffusion Distribution

$$\hat{z}_{0|t} = \frac{1}{\sqrt{\bar{\alpha}_t}}\left(z_t - \sqrt{1-\bar{\alpha}_t}\,\epsilon_\theta(z_t,t)\right)$$

$$\mu_{\theta,t} = \sqrt{\bar{\alpha}_{t-1}}\hat{z}_{0|t} + \sqrt{1-\bar{\alpha}_{t-1}-\sigma_t^2}\,\epsilon_\theta(z_t,t)$$

$$z_{t-1} = \mu_{\theta,t} + \sigma_t\left(\frac{\sin[(1-\lambda)\beta]}{\sin\beta}\epsilon_t - \frac{\sin(\lambda\beta)}{\sin\beta}\cdot\frac{\|\epsilon_t\|_F\nabla_{z_t}\mathcal{L}}{\|\nabla_{z_t}\mathcal{L}\|_F}\right)$$

where $\beta = \angle(\epsilon_t, -\nabla_{z_t}\mathcal{L})$

Implementation Details

- **Dataset:** MoisesDB and MusicCaps 100 music tracks.
- **Preprocessing:**
  - 5-second non-overlapping segments per track.
  - Audio sampled at 16 kHz.
  - Log-mel spectrograms: 1024 FFT, hop size 160, 64 mel bins.
- **Backbone:** AudioLDM2 (frozen)
- **Sampling:** DDIM 500 steps, null-text conditioning.
- **Metrics:** LSD, FAD.

Music Restoration Tasks

$$\mathcal{L}(x,y) = \|\mathcal{A}(x) - y\|_F$$

Task-specific operator Degraded music

Dataset	Inpainting		Super-Resolution		Dereverberation		Phase Retrieval	
	LSD ↓	FAD ↓	LSD ↓	FAD ↓	LSD ↓	FAD ↓	LSD ↓	FAD ↓
MoisesDB								
DPS [17]	0.7960	0.4847	1.1019	0.5794	1.1985	0.6482	**0.6973**	0.5325
MPGD [18]	1.7190	0.6108	1.7423	0.6096	1.7386	0.6011	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**	0.7056	**0.4666**
MusicCaps								
DPS [17]	0.9026	0.5185	1.0453	0.4727	1.0449	0.5000	**0.7388**	0.5683
MPGD [18]	1.2734	0.5153	1.2864	0.5151	1.2815	0.5115	1.2632	0.5142
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**	0.8051	**0.4430**

Qualitative Results: Inpainting

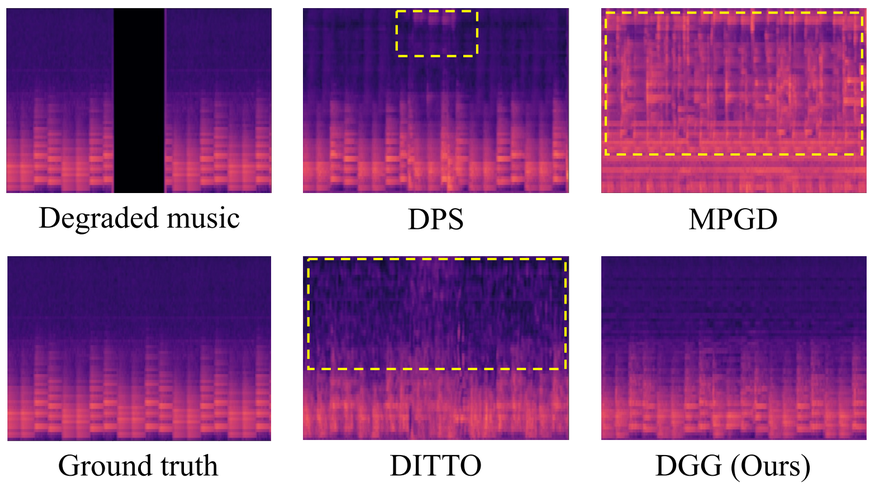


Figure. Qualitative comparison of mel-spectrograms for the inpainting task. Dashed boxes highlight regions with noticeable generation or reconstruction artifacts.

Key Contributions

- Prevent latent drift by updating latents along diffusion geodesics using spherical linear interpolation.
- Our method supports multiple music restoration tasks within a unified framework: inpainting missing audio parts, audio super-resolution, dereverberation, and phase retrieval, all in a zero-shot manner, i.e., no task-specific retraining.
- Empirically, DGG consistently outperforms state-of-the-art *training-free* baselines across tasks and datasets, showing lower spectral distortion (LSD), better perceptual audio quality (FAD), and comparable inference speed.