DiffMusic: A Zero-shot Diffusion-Based Framework for Music Inverse Problem

Is Training All You Need?

Jia-Wei Liao

Pin-Chi Pan

Sheng-Ping Yang

National Taiwan University

https://github.com/jwliao1209/DiffMusic

Outline

- Introduction
- Proposed Method
- Experiments and Demo
- Summary

Introduction

- **Motivation**
  - Enhancing music processing quality and application diversity.
  - Designing the method without training or fine-tuning under limited computational resources.

- **Goal:** Developing a **zero-shot diffusion-based framework** to address music related **inverse problems**, such as music inpainting, super-resolution, phase retrieval, source separation, and music dereverberation.

Innovation Contribution

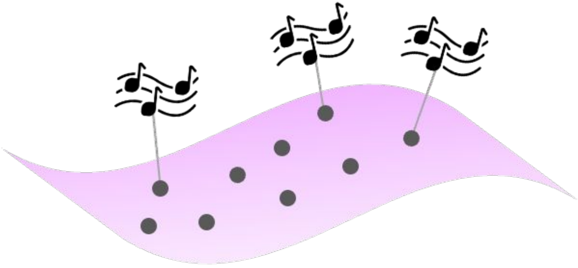
- We propose a **training-free** diffusion-based framework capable of addressing music inverse problems within 1 minute.
- We introduce a **Vocoder-Mel Constraint (VMC)** to enhance the quality of generated music.
- Our pipeline enables **iterative refinement** through sampling processes with a pretrained model (e.g. AudioLDM2, MusicLDM) and supports **plug-and-play** adaptability, expanding its applications in the music domain.

Music Inverse Problems (IP)

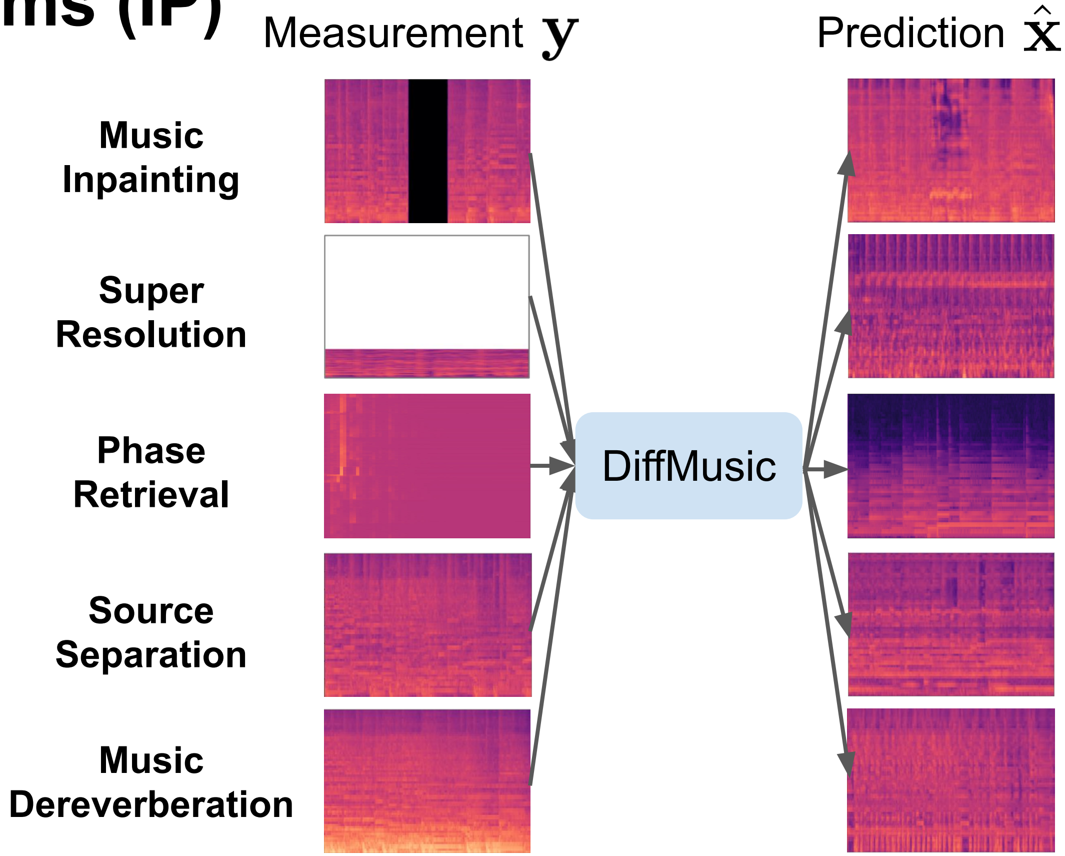
Operator

$$\hat{\mathbf{x}} = \operatorname*{argmin}_{\mathbf{x}\in\mathcal{M}} \|\mathbf{A}\mathbf{x} - \mathbf{y}\|_2^2$$

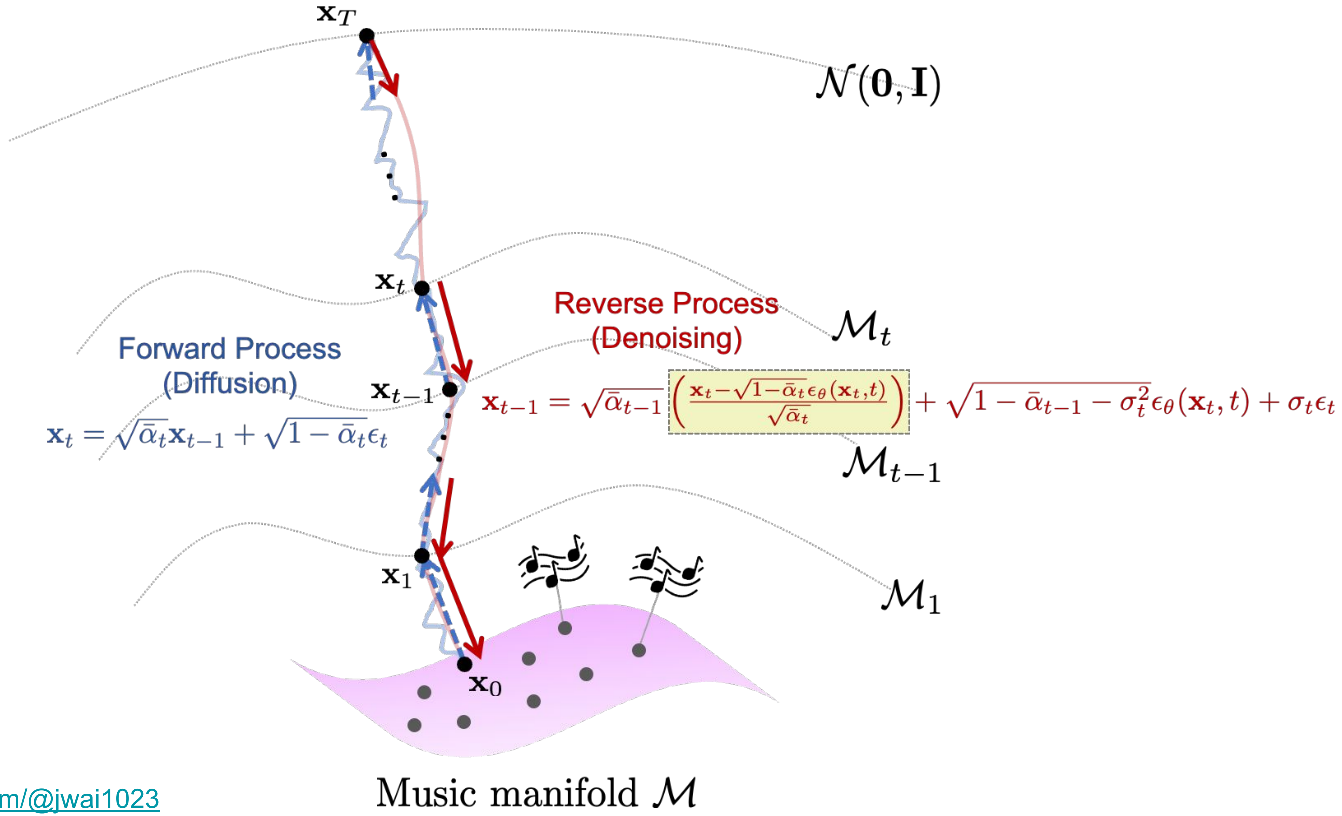
Measurement



Music manifold $\mathcal{M}$

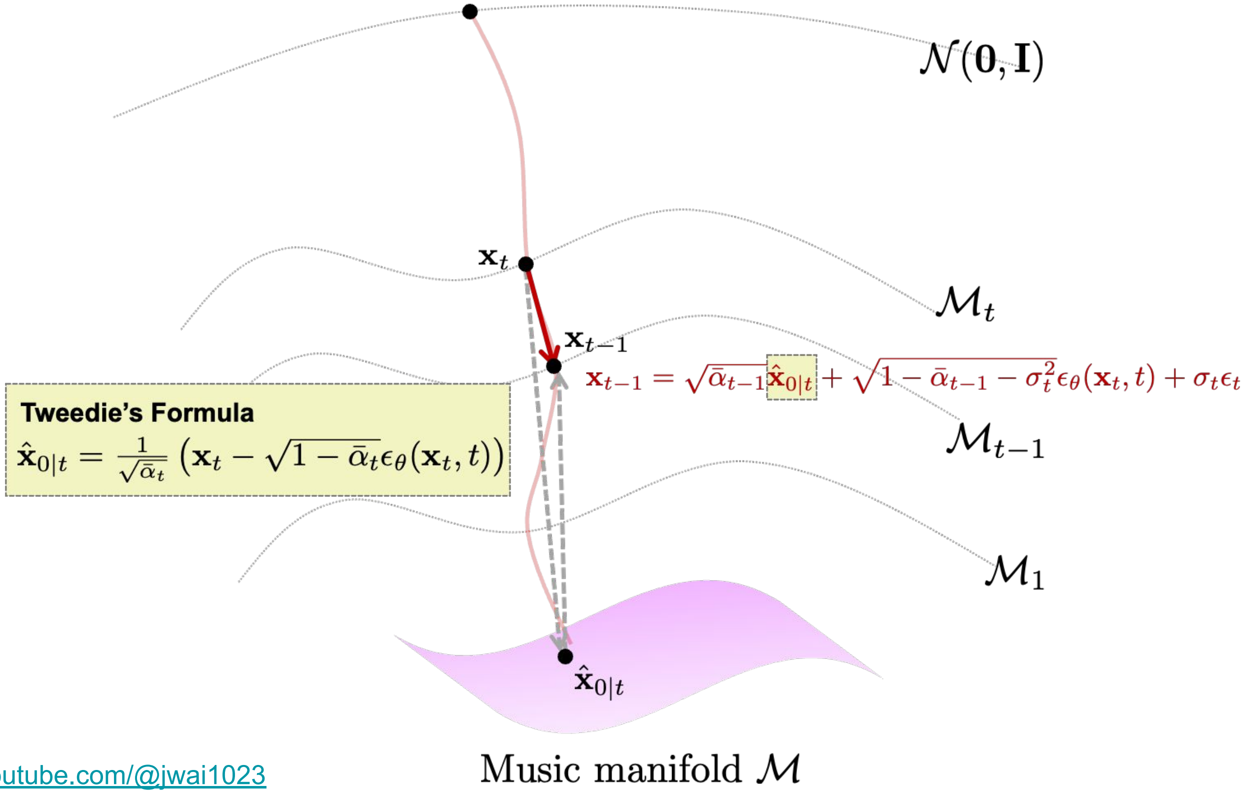


Sampling Process



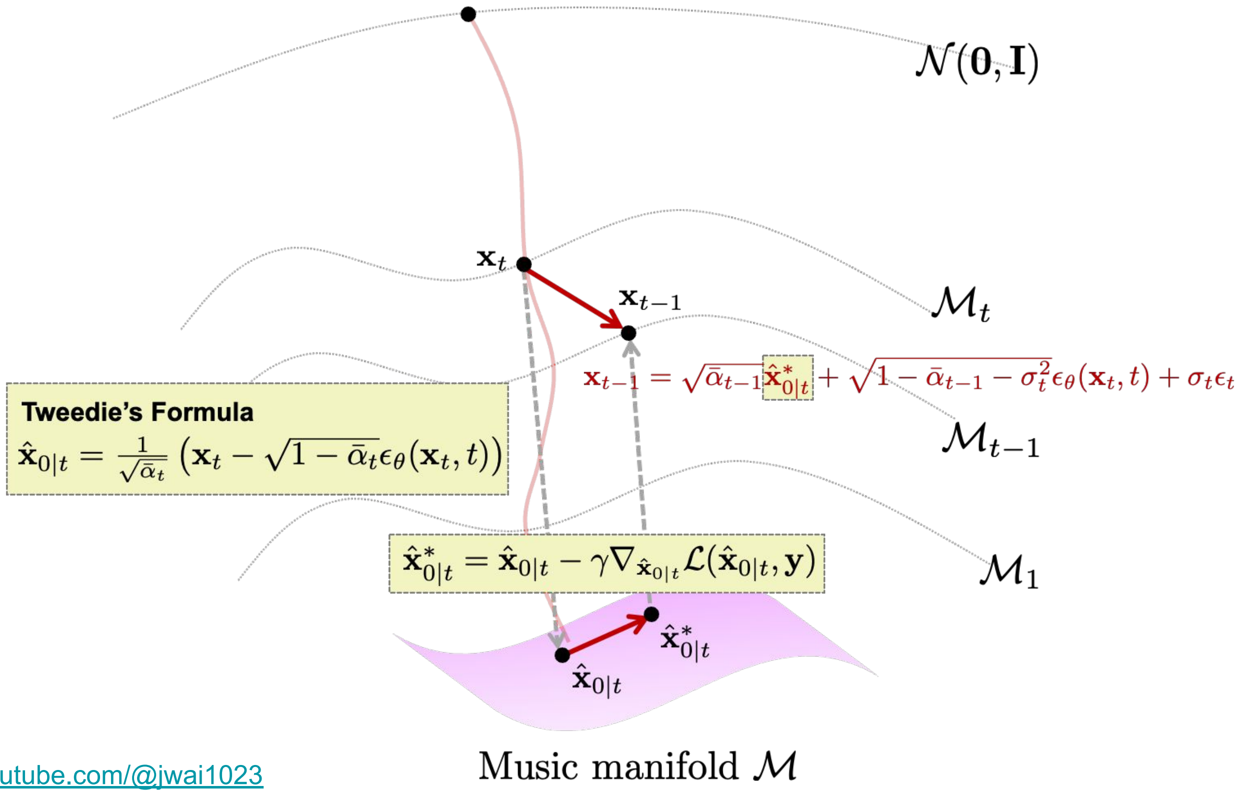
https://www.youtube.com/@jwai1023

Sampling Process



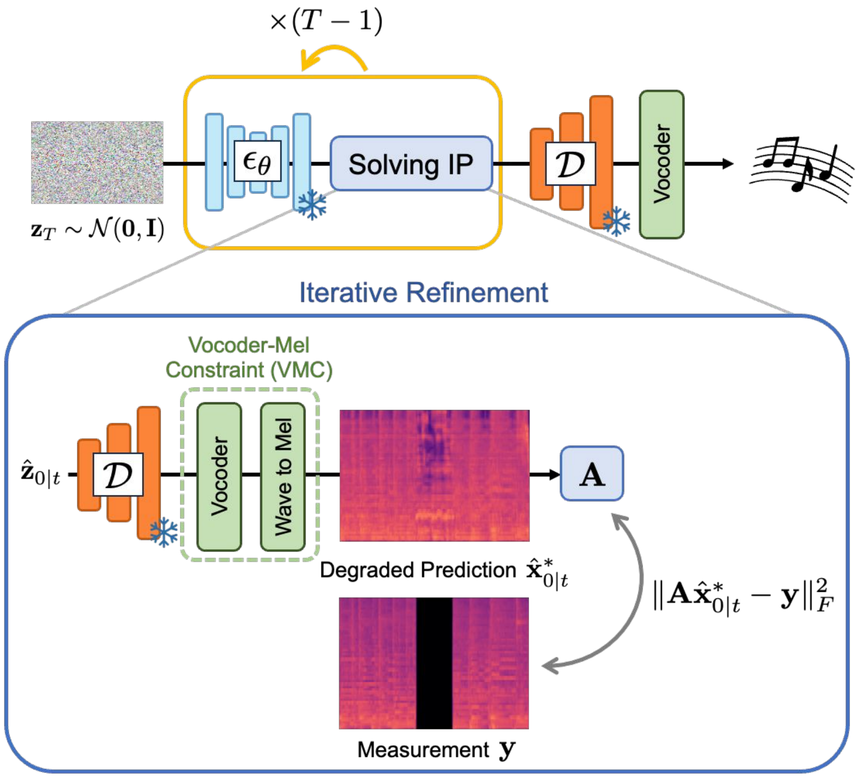
https://www.youtube.com/@jwai1023

Sampling with Iterative Refinement



https://www.youtube.com/@jwai1023

Proposed Pipeline



Key Design:

Decoder: Project latent to manifold tangent space

VMC: Preserve the quality of audio

Algorithm 1 DiffMusic

Input: Measurement $\mathbf{y}$, UNet $\epsilon_\theta(\cdot)$, VAE decoder $\mathcal{D}(\cdot)$, wave to mel spectrogram transformation T, vocoder $\mathcal{V}(\cdot)$, sequence of noise schedule $\{\bar{\alpha}_t\}_{t=1}^T$, learning rate $\gamma > 0$.

$\mathbf{z}_T \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$.

for $t = T$ to 1 **do**

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

  $\hat{\mathbf{x}}_{0|t} \leftarrow (T \circ \mathcal{V} \circ \mathcal{D})(\hat{\mathbf{z}}_{0|t})$.

  $\hat{\mathbf{z}}^*_{0|t} \leftarrow \hat{\mathbf{z}}_{0|t} - \gamma \nabla_{\hat{\mathbf{z}}_{0|t}} \|\mathbf{A}\hat{\mathbf{x}}_{0|t} - \mathbf{y}\|_F^2$.

  $\hat{\epsilon}_t \leftarrow \frac{\mathbf{z}_t - \sqrt{\bar{\alpha}_t}\hat{\mathbf{z}}^*_{0|t}}{\sqrt{1-\bar{\alpha}_t}}$.

  $\mathbf{z}_{t-1} \leftarrow \sqrt{\bar{\alpha}_{t-1}}\hat{\mathbf{z}}^*_{0|t} + \sqrt{1-\bar{\alpha}_{t-1}}\hat{\epsilon}_t$.

end for

return $(\mathcal{V} \circ \mathcal{D})(\mathbf{z}_0)$.

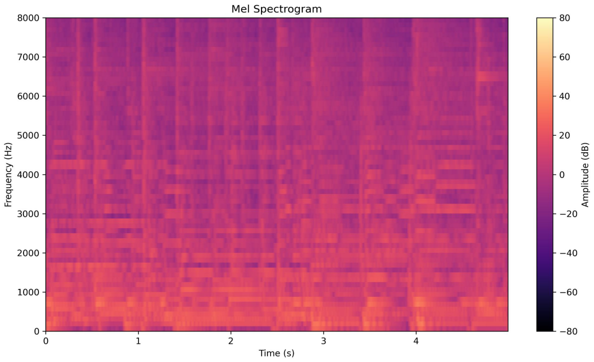
Inverse Problem: Music Inpainting

Filling in missing or damaged parts of a musical piece to restore continuity and maintain its original style.

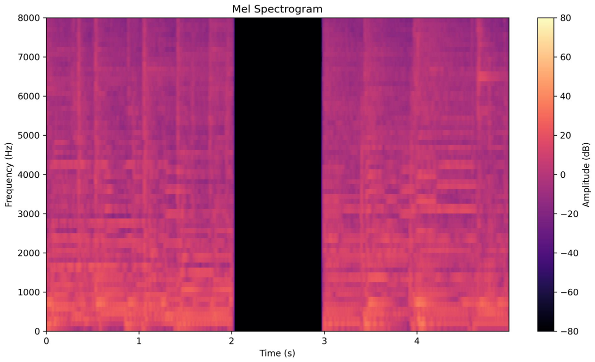
$$\mathcal{L} = \|\mathbf{A} \odot \mathbf{x} - \mathbf{y}\|_F^2$$

$$\mathbf{A}_{f,t} = \begin{cases} 0, & \text{if } \in [t_{\text{start}}, t_{\text{end}}], \forall f \\ 1, & \text{otherwise.} \end{cases}$$

Original Audio (***x***)



Measurement (***y***)



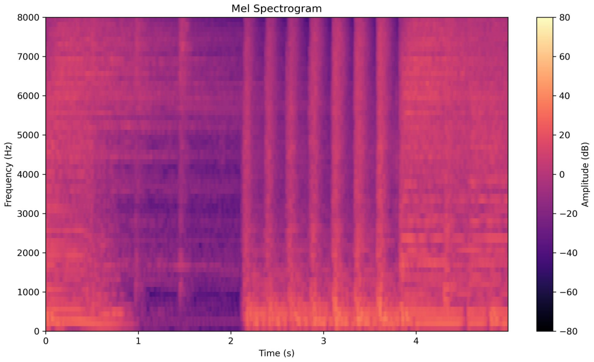
Inverse Problem: Super-Resolution

Enhancing the quality of audio signals by reconstructing high-resolution audio from low-resolution input.

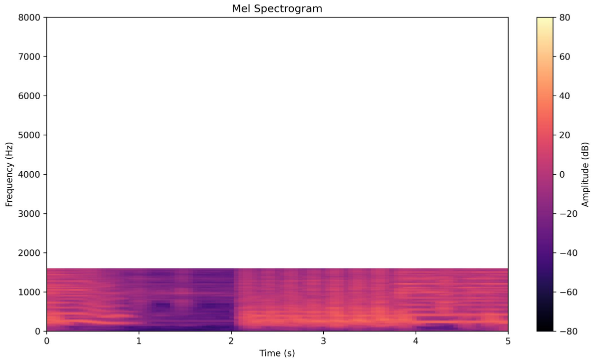
$$\mathcal{L} = \|\mathbf{R}(\mathbf{x}) - \mathbf{y}\|_F^2$$

Resample

Original Audio (***x***)



Measurement (***y***)

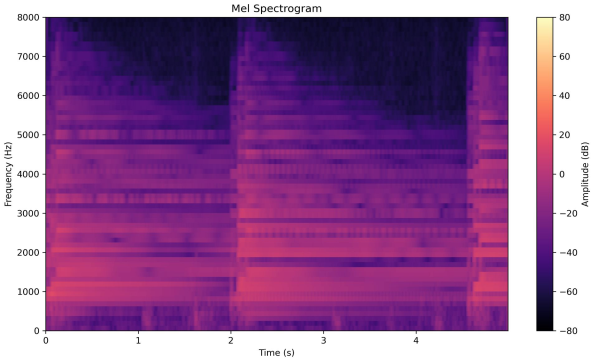


Inverse Problem: Phase Retrieval

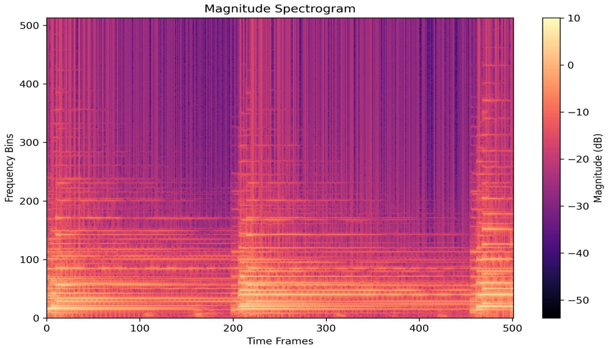
Reconstructing a complete audio signal by estimating its phase from spectral amplitude.

$$\mathcal{L} = \| |\mathrm{STFT}(\mathbf{x})| - \mathbf{y} \|_F^2$$

Original Audio (***x***)



Measurement (***y***)

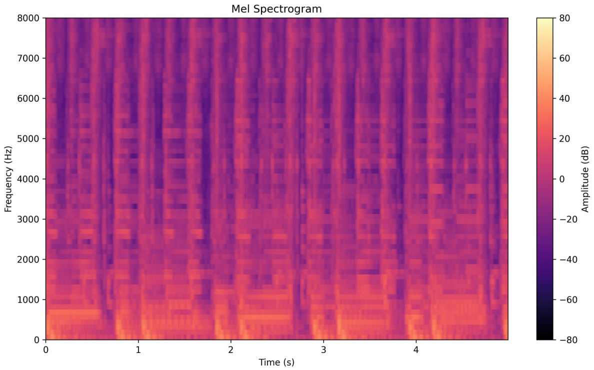


Inverse Problem: Source Separation

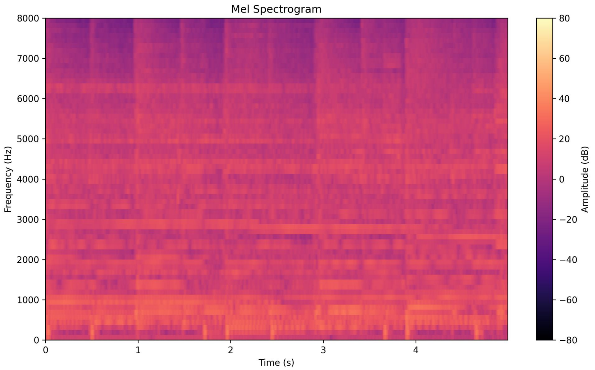
Removing noise or isolating specific audio elements (e.g., vocals, instruments) from a mixed signal.

$$\mathcal{L} = \|w\mathbf{x}_1 + (1-w)\mathbf{x}_2 - \mathbf{y}\|_F^2$$

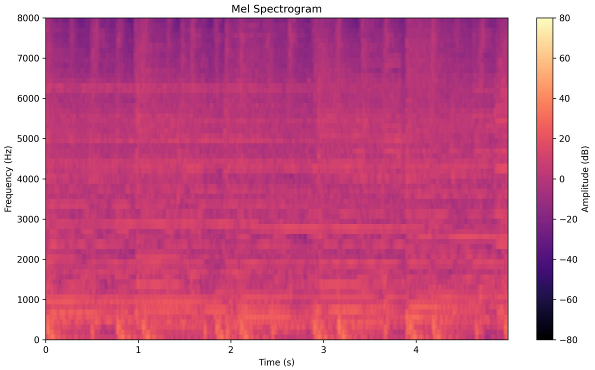
Original Audio ($\boldsymbol{x}_1$)



Residual Audio ($\boldsymbol{x}_2$)



Measurement ($\boldsymbol{y}$)



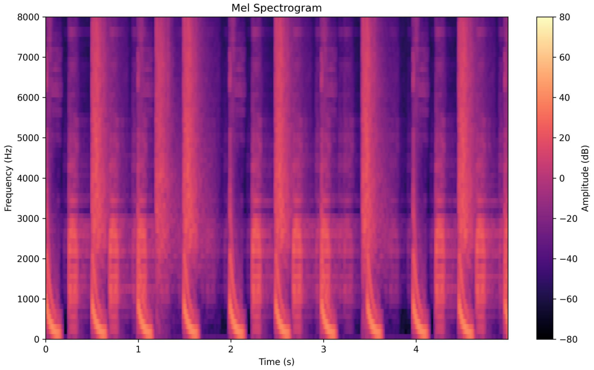
Inverse Problem: Music Dereverberation

Removing reverberation effects to recover a clean audio signal, free from echoes caused by room reflections.

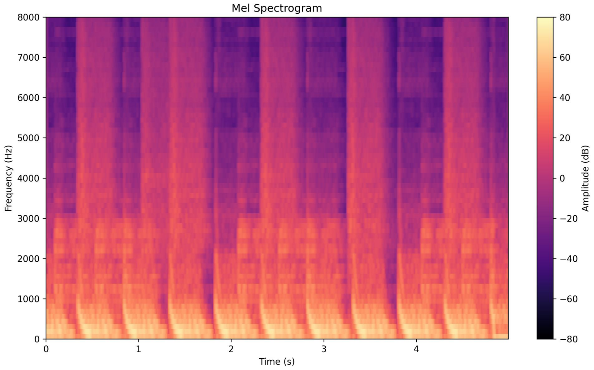
$$\mathcal{L} = \|\mathbf{h} * \mathbf{x} - \mathbf{y}\|_F^2$$

Reverberation Impulse Response

Original Audio (***x***)



Measurement (***y***)



Experiments

Dataset: Musdb18 100 songs

Model: AudioLDM2, MusicLDM

- **LSD:** Log Spectral Distance

$$\frac{1}{N}\sum_{n=1}^{N}\sqrt{\frac{1}{K}\sum_{k=1}^{K}\left(\log|X_{\mathrm{rec}}(n,k)|-\log|X_{\mathrm{gt}}(n,k)|\right)^2}$$

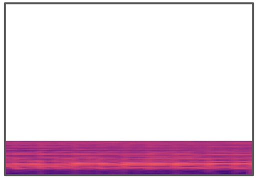
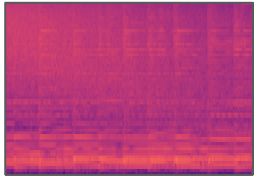
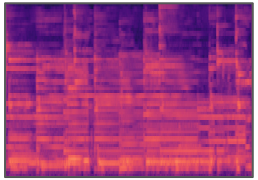
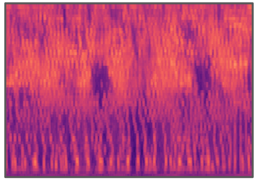
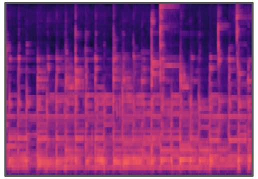
- **FAD:** Fréchet Audio Distance

$$\|\mu_{\mathrm{rec}}-\mu_{\mathrm{gt}}\|^2+\mathrm{Tr}\left(\Sigma_{\mathrm{rec}}+\Sigma_{\mathrm{gt}}-2\left(\Sigma_{\mathrm{rec}}\Sigma_{\mathrm{gt}}\right)^{1/2}\right)$$

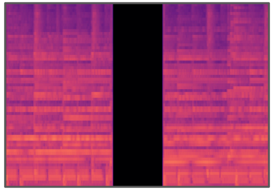
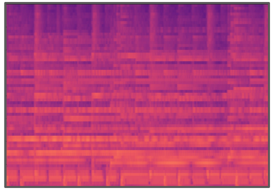
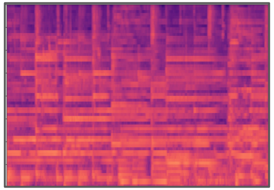
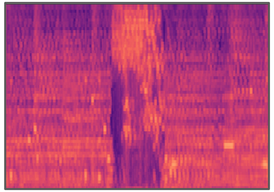
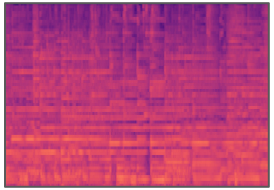
Inverse Problem	Methods	AudioLDM2 LSD ↓	AudioLDM2 FAD ↓	MusicLDM LSD ↓	MusicLDM FAD ↓
Music Inpainting	DPS	**0.6207**	6.4334	**0.6318**	5.1730
	DSG	0.7699	13.8478	0.7735	14.0801
	DiffMusic (Our)	0.6341	**4.9896**	0.6367	**4.3202**
Super Resolution	DPS	0.9815	9.2984	**0.9351**	7.9806
	DSG	1.3427	15.0559	1.2783	17.1117
	DiffMusic (Our)	**0.9678**	**8.9296**	0.9778	**5.8756**
Phase Retrieval	DPS	**0.8180**	7.7626	**0.7653**	6.7907
	DSG	0.8258	14.6598	0.8873	16.4876
	DiffMusic (Our)	0.8323	**6.2551**	0.8939	**4.6492**
Source Separation	DPS	0.9350	7.9542	0.9603	5.8537
	DSG	0.8241	14.0942	0.9120	16.9260
	DiffMusic (Our)	**0.8293**	**6.2551**	**0.9334**	**4.9374**
Music Dereverberation	DPS	0.6837	7.8759	0.7536	5.7646
	DSG	0.7560	13.8926	0.8308	16.7926
	DiffMusic (Our)	**0.6604**	**7.1838**	**0.6788**	**4.8319**

[DPS] Diffusion Posterior Sampling for General Noisy Inverse Problems
[DSG] Guidance with Spherical Gaussian Constraint for Conditional Diffusion

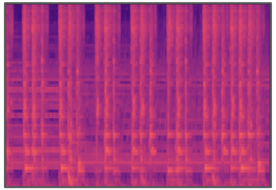
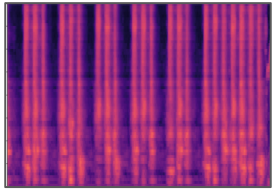
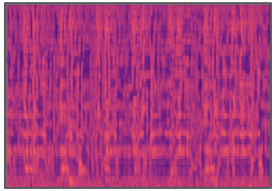
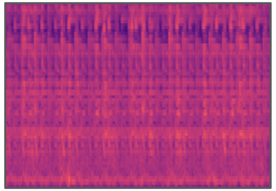
Demo Case: Super Resolution

	Input	Label	DPS	DSG	DiffMusic
Audio Sample					
Mel Spectrogram					

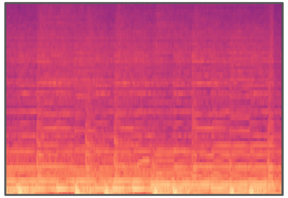
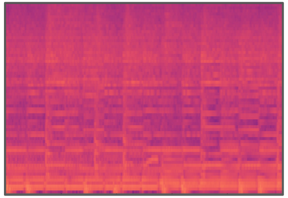
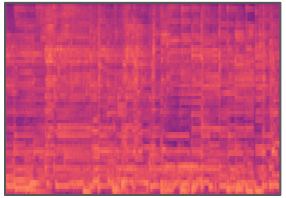
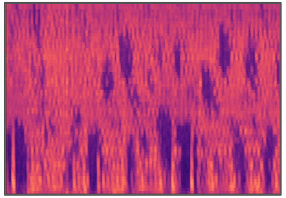
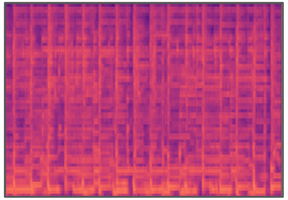
Demo Case: Music Inpainting

	Input	Label	DPS	DSG	DiffMusic
Audio Sample					
Mel Spectrogram					

Demo Case: Phase Retrieval

	Label	DPS	DSG	DiffMusic
Audio Sample				
Mel Spectrogram				

Demo Case: Dereverberation

	Input	Label	DPS	DSG	DiffMusic
Audio Sample					
Mel Spectrogram					

Summary

1. We propose DiffMusic, a zero-shot diffusion-based framework, designed to solve various music inverse problems.

2. Leverages pretrained models for zero-shot conditional generation, provide 5 operation to enable flexible music processing without extensive fine-tuning.

3. Experimental results show flexible performance across different tasks, highlighting DiffMusic’s potential in enhancing music restoration and multi-task generation.

Thank you