Project Page

100+ generated music samples to be listened!

Efficient Neural Music Generation

NeurIPS 2023 Poster

Max W. Y. Lam, Qiao Tian, Tang Li, Zongyu Yin, Siyuan Feng, Ming Tu, Yuliang Ji, Rui Xia, Mingbo Ma, Xuchen Song, Jitong Chen, Yuping Wang, Yuxuan Wang

Speech, Audio & Music Intelligence (SAMI), ByteDance

Project Page: https://efficient-melody.github.io/

Introduction

- Music audio generation has recently been advanced by the audio language modeling (LM) approach (Borsos et al., 2022; Agostinelli et al., 2023).
- The state-of-the-art (SOTA) MusicLM employs a two-stage modeling framework: **semantic modeling** followed by **acoustic modeling**.
- Acoustic modeling in MusicLM entails predicting multiple RVQ tokens, thus defines separately trained **coarse and fine acoustic LMs**.
- MusicLM requires sequentially processing through 3 LMs for generation, making it computationally expensive and prohibitive for a long generation.
- Efficient music generation with a quality on par with MusicLM remains a significant challenge.
- We propose **MeLoDy** (**M** for music; **L** for LM; **D** for diffusion), an LM-guided diffusion model that generates music audios of state-of-the-art quality and reduces **95.7% to 99.6%** forward passes in MusicLM for sampling 10s to 30s music.

Background

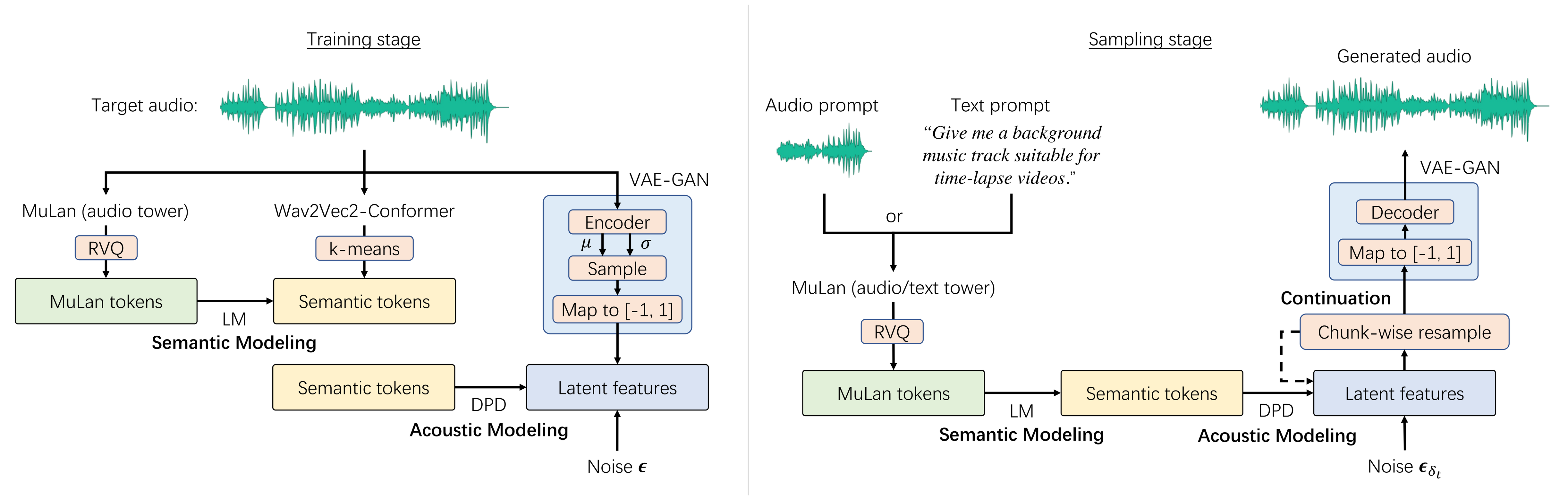
Conventional text-to-music generation models:

Model	Data	AC	FR	VT	MP
Moûsai (2023)	2.5kh	✓	✓	✗	✗
MusicLM (2023)	280kh	✓	✗	✓	✗
Noise2Music (2023)	340kh	✗	✗	✓	✗
MusicGen (Parallel)	20kh	✓	✓	✓	✗
MeLoDy (Ours)	257kh	✓	✓	✓	✓

- **AC**: supports audio continuation
- **FR**: is faster than real-time on a V100 GPU
- **VT**: was tested with a variety of text prompts
- **MP**: was evaluated by music producers

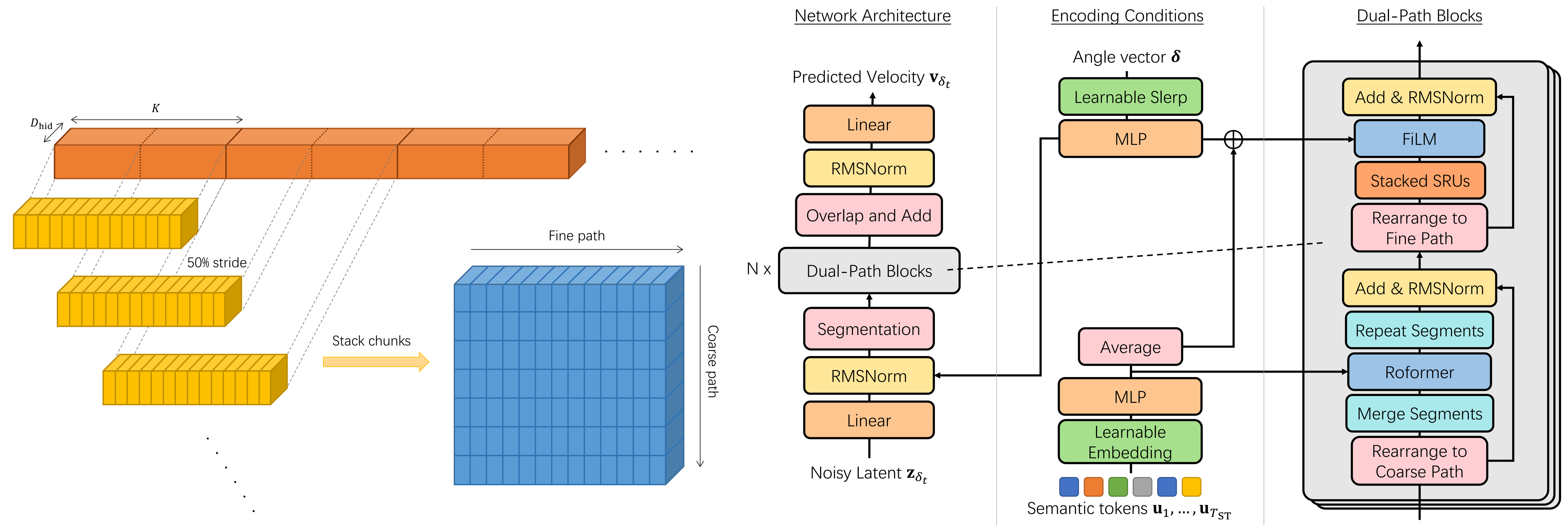
MeLoDy is the first large-scale trained model that satisfies both **AC**, **FR**, **VT** and **MP**.

The proposed MeLoDy pipeline



- The proposed MeLoDy pipeline is inherited from the MusicLM framework, but is much more efficient with the **coarse-and-fine acoustics being simultaneously modeled** in one DPD model.
- A critical difference between acoustic LMs and DPD is the definition of auto-encoder: **Neural codec v.s. Audio VAE-GAN** (similar to SD); **Discrete tokens v.s. Continuous latents**.

Dual-Path Diffusion (DPD)



- DPD is a variant of latent diffusion models (LDMs) that operates on angular space: For angle $\delta \in [0, \pi/2]$, $\mathbf{z}_\delta = \cos(\delta)\mathbf{z}_0 + \sin(\delta)\boldsymbol{\epsilon}$, $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. ($\mathbf{z}_\delta$ gets noisier as δ increases to $\pi/2$).
- To learn coarse and fine acoustics in one model, we design a dual-path modeling scheme based on **(i) segmentation (left)**, and **(ii) alternating coarse and fine paths (right)** in each DPD block.
- Effective conditioning approaches are proposed for DPD: **(i) learnable Slerp for δ-encoding**, and **(ii) coarse-path cross-attention & fine-path FiLM conditioning**.

Results

1) Speed and quality analysis:

Steps	Speed (CPU)	Speed (GPU)	FAD
5	1472Hz (0.06×)	181.1kHz (7.5×)	7.23
10	893Hz (0.04×)	104.8kHz (4.4×)	5.93
20	498Hz (0.02×)	56.9kHz (2.4×)	5.41

2) Pair-wise compare to MusicLM:

Model	Musicality	Quality	Text Corr.
MusicLM	**54.1%**	46.5%	**54.8%**
MeLoDy	45.9%	**53.5%**	45.2%

3) Pair-wise compare to Noise2Music:

Model	Musicality	Quality	Text Corr.
Noise2Music	**55.5%**	43.6%	**57.2%**
MeLoDy	44.5%	**56.4%**	42.8%

4) Ablation on network architecture:

Network	Velocity MSE	SI-SNRi
UNet-1D	0.13	5.33
UNet-2D	0.15	4.96
DPD	**0.12**	**6.15**

5) Ablation on angle schedule:

Angle schedule	Steps	FAD
Uniform: $\omega_t = \frac{\pi}{2T}$	10	8.52
	20	6.31
Ours: $\omega_t = \frac{\pi}{6T} + \frac{2\pi t}{3T(T+1)}$	10	**5.93**
	20	**5.41**

Broader Impact

- MeLoDy practically facilitates content creators to express their creative pursuits with text prompts.
- In the light of efficient sampling, MeLoDy also enables an interactive creation process to take human feedback into account.