MusicLDM: Enhancing Novelty in Text-to-Music Generation Using Beat-Synchronous Mixup Strategies

Ke Chen*[1] Yusong Wu*[2] Haohe Liu*[3] Marianna Nezhurina[4]
Taylor Berg-Kirkpatrick[1] Shlomo Dubnov[1]

[1]University of California San Diego (UCSD)
[2]Mila, Quebec Artificial Intelligence Institute, University of Montreal
[3]Centre for Vision Speech and Signal Processing, University of Surrey
[4]LAION

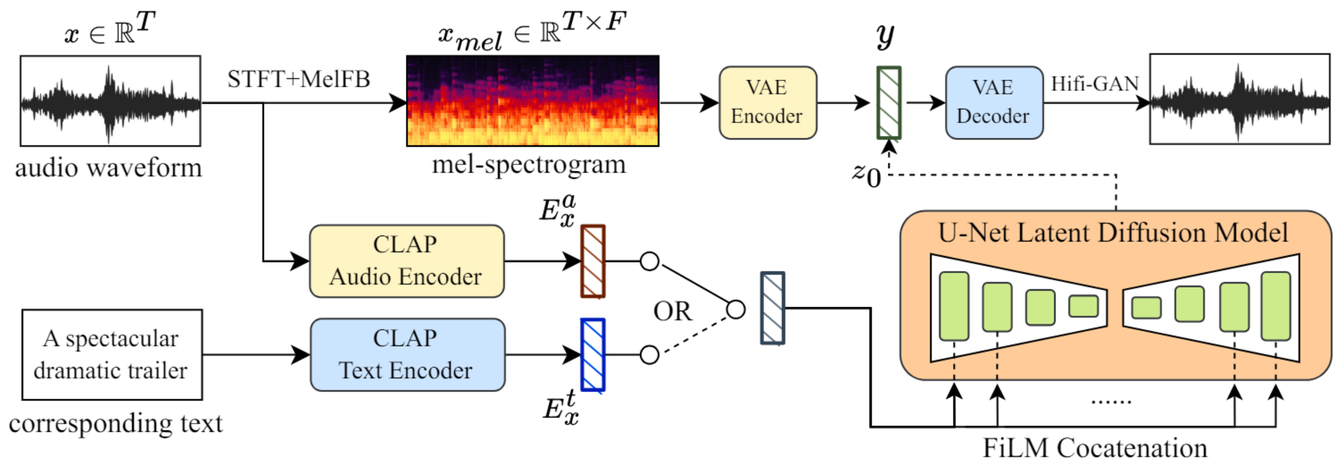
Introduction

MusicLDM is a text-to-music generation model focusing on enhancing the novelty, namely avoiding the plagiarism in the training data and embracing more data diversity, of the generation results.

The paper contains:

- A text-to-music generator based on the latent diffusion model (LDM)
- A training process with the latent mix-up strategy to increase the diversity of training data
- A comprehensive evaluation in both music generation quality, audio similarity and relevance

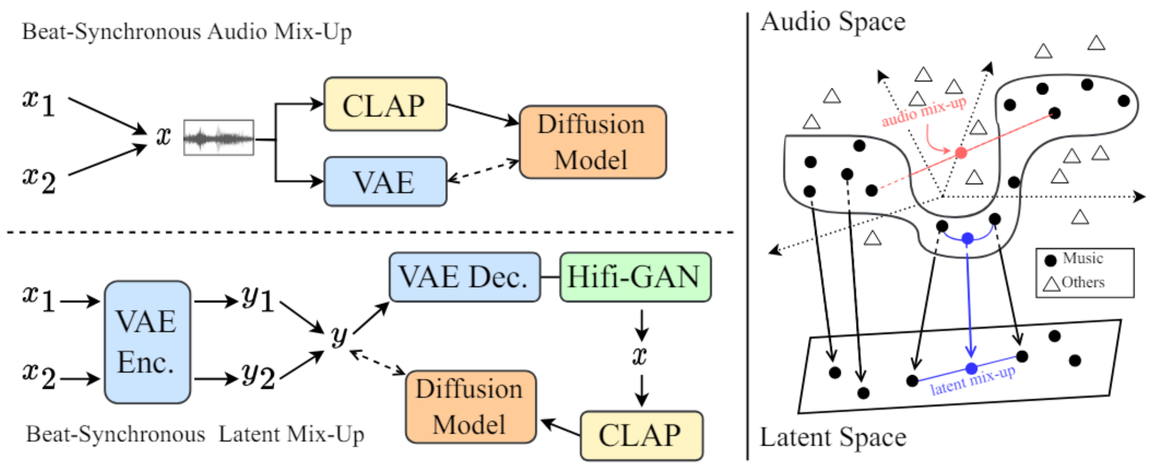
Model Architecture



Generation Quality Evaluation

Model	FD_{pann} ↓	FD_{vgg} ↓	Inception Score ↑	KL Div. ↓
Riffusion [11]	68.95	10.77	1.34	5.00
MuBERT [8]	31.70	19.04	1.51	4.69
AudioLDM (w/. original CLAP) [6]	38.92	3.08	1.67	3.65
Moûsai [24]	30.73	10.59	1.50	3.88
MusicGen* [10]	**25.19**	**2.17**	**1.82**	**3.10**
MusicLDM	26.67	2.40	**1.81**	3.80
MusicLDM (Only TA-Training)	32.40	2.51	1.49	3.96
MusicLDM w/. mixup	30.15	2.84	1.51	3.74
MusicLDM w/. BAM	28.54	**2.26**	1.56	3.50
MusicLDM w/. BLM	**24.95**	2.31	1.79	**3.40**
MusicLDM w/. Text-Finetune	27.81	1.75	1.76	3.60
MusicLDM w/. BAM & Text-Finetune	28.22	1.81	1.61	3.61
MusicLDM w/. BLM & Text-Finetune	**26.34**	**1.68**	**1.82**	**3.47**

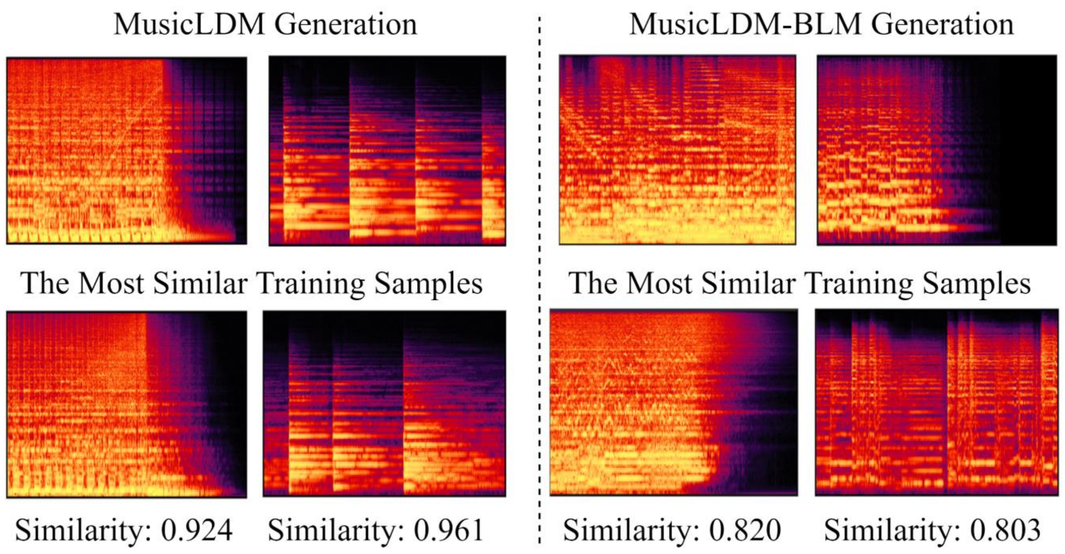
Beat-Synchronous Audio Mix-up



Similarity & Relevance Evaluation

Model	Objective Metrics			Subjective Listening Test		
	T-A Similarity	SIM_{AA}@90	SIM_{AA}@95	Quality	Relevance	Musicality
MusicLDM	**0.281**	0.430	0.047	1.98	2.17	**2.19**
MusicLDM-mixup	0.234	**0.391**	0.028	—	—	—
MusicLDM-BAM	0.266	0.402	0.027	2.04	2.21	2.01
MusicLDM-BLM	0.268	0.401	**0.020**	**2.13**	**2.31**	2.07

Visualization



QR Codes

Demo

Paper

Code