MusicGen

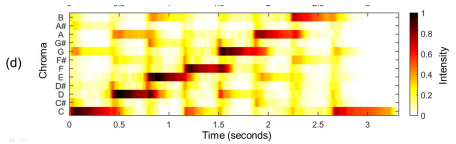
Simple and Controllable Music Generation

Исследователь: Писцов Георгий 22.01.24

О чем сегодня поговорим

- MusicGen
- Riffusion
- MusicLM
- Mousai
- Noise2Music
- Выводы

MusicGen



Melody chromogram

Encoded text

argmax over time

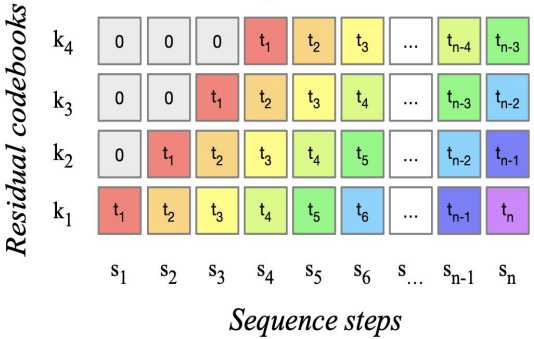
Encoder

RVQ

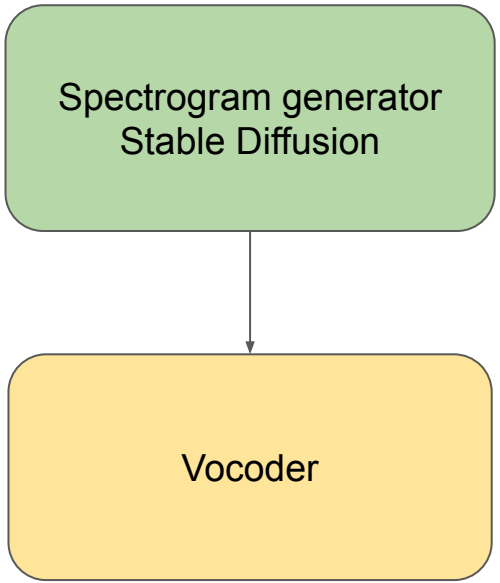
Codebook projection

Multi-layer Transformer Decoder

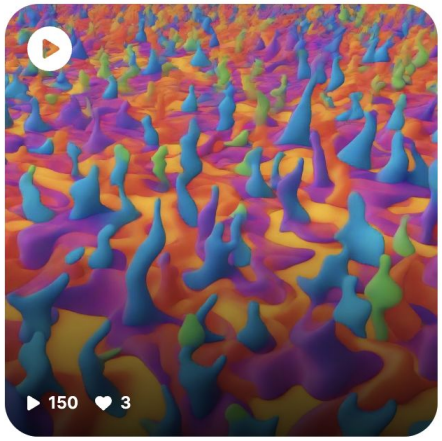
Delay Pattern



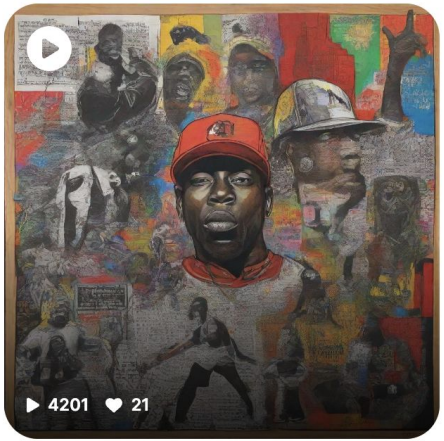
Riffusion



- Fine-tuned checkpoint from Stable Diffusion
- No paper, pet-project

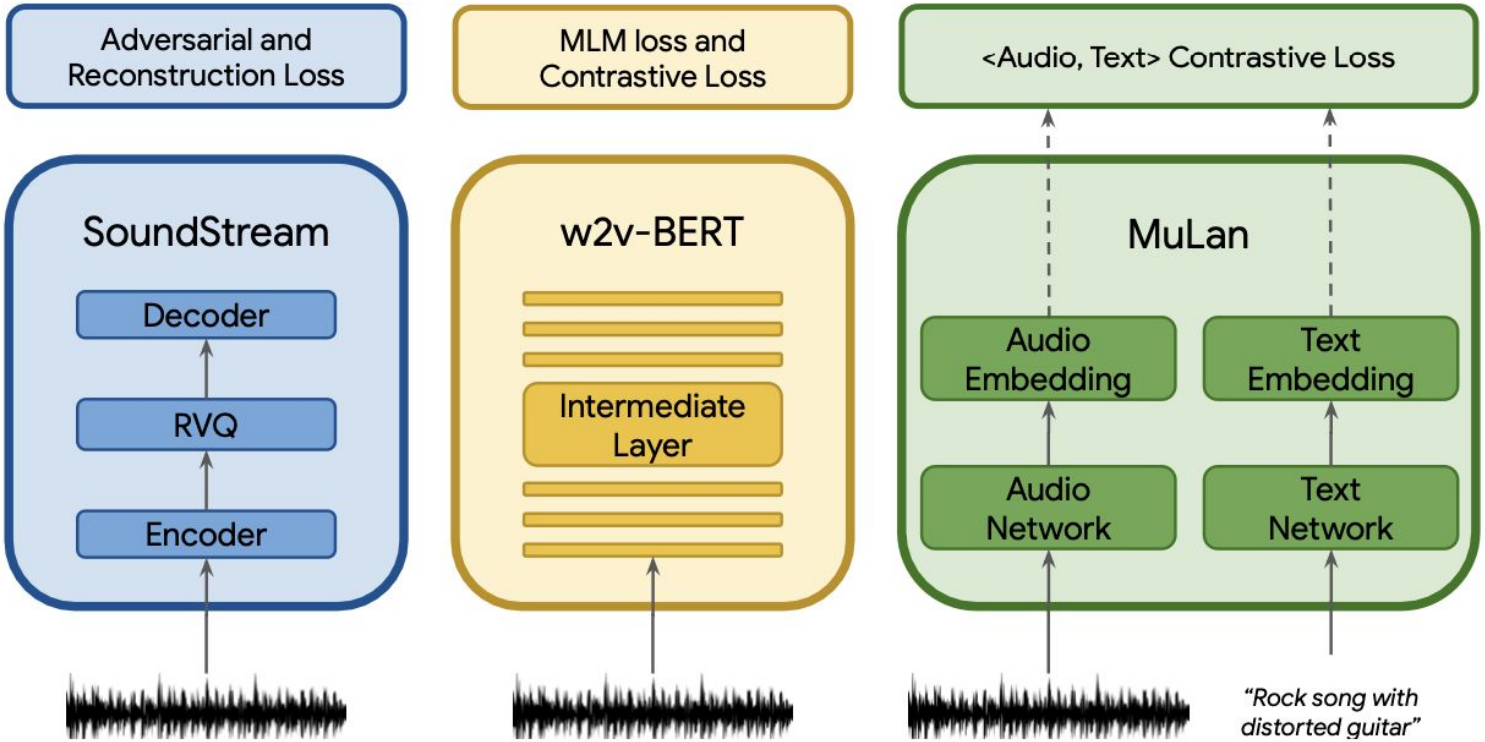


L **Poop Dance Revolution**
Landon Walton • 1d ago



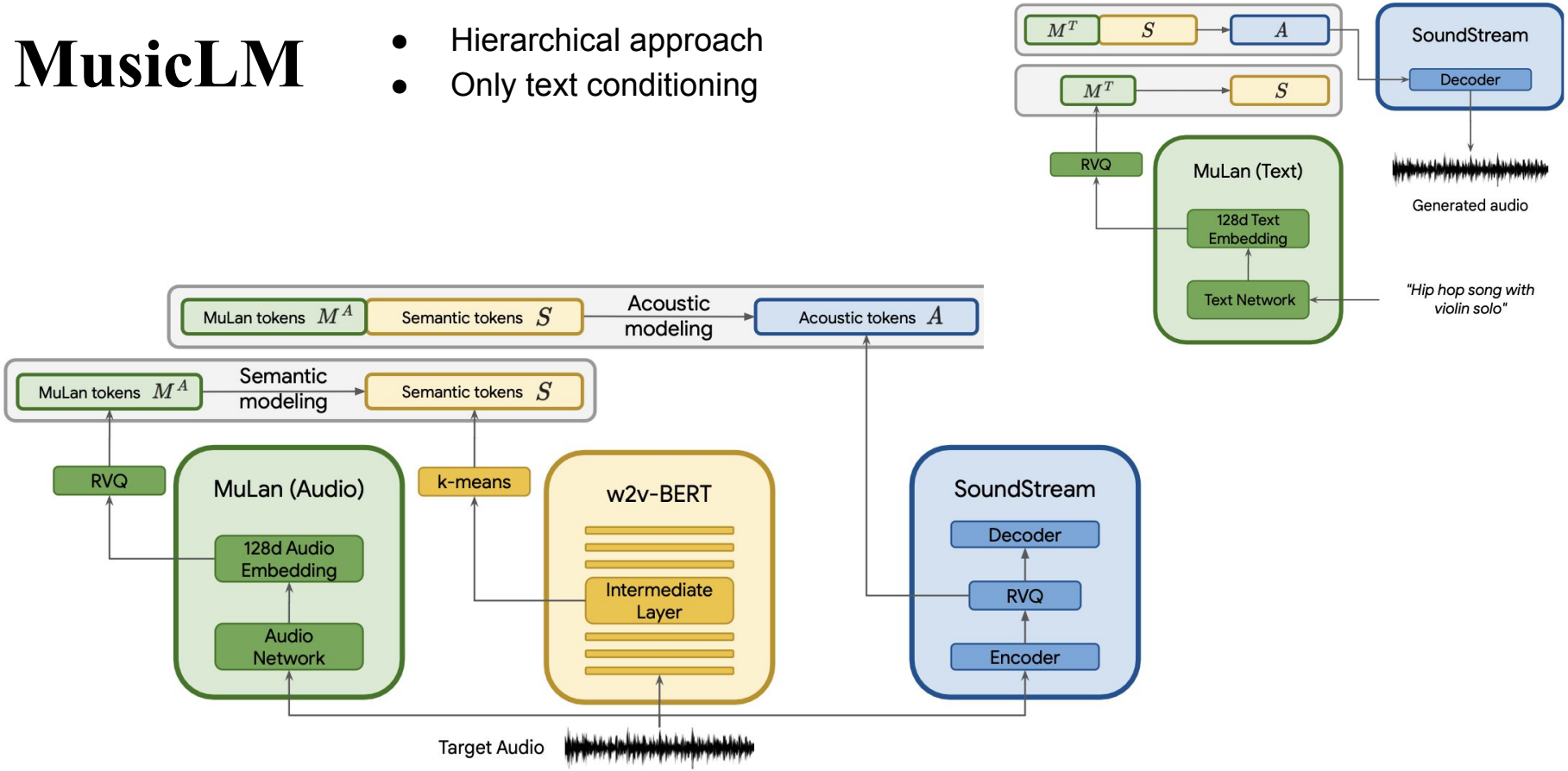
A **Check the Rap, Know the Difference**
Amanda Martinez • 1d ago

MusicLM

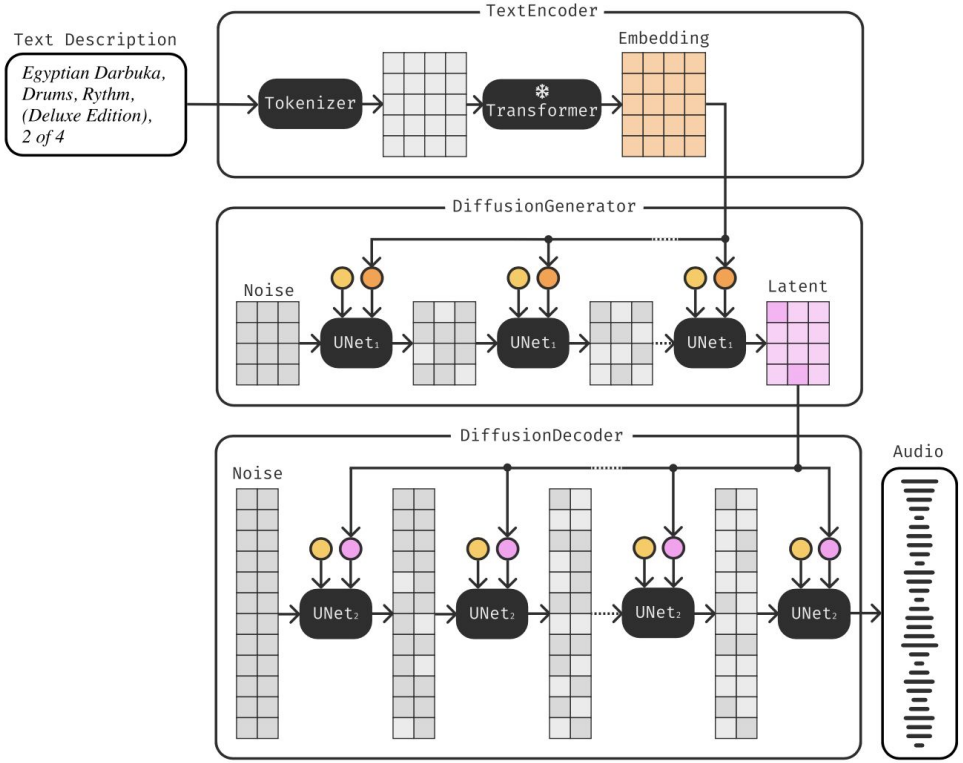

Adversarial and Reconstruction Loss
MLM loss and Contrastive Loss
<Audio, Text> Contrastive Loss
SoundStream
Decoder
RVQ
Encoder
w2v-BERT
Intermediate Layer
MuLan
Audio Embedding
Text Embedding
Audio Network
Text Network
"Rock song with distorted guitar"

MusicLM

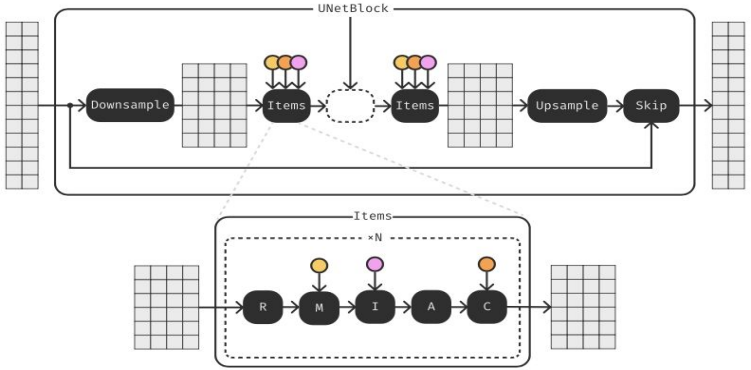
- Hierarchical approach
- Only text conditioning



Mousai: Efficient Text-to-Music Diffusion

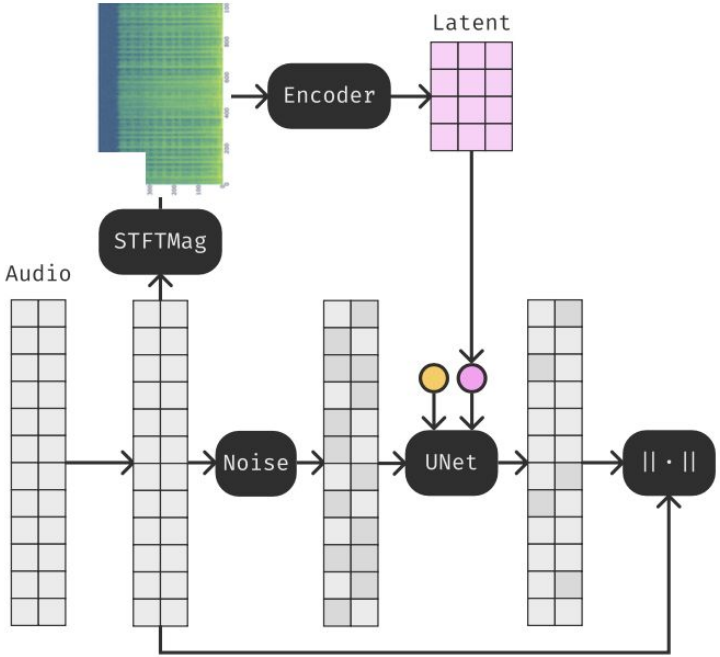


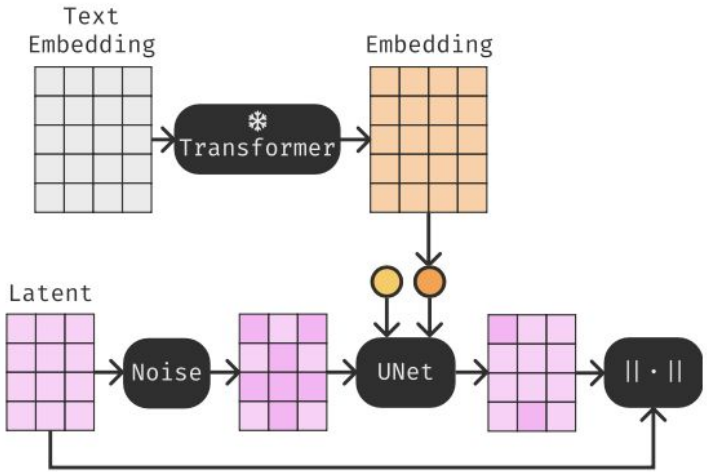
- 2-stage cascading diffusion approach; only text-conditioning



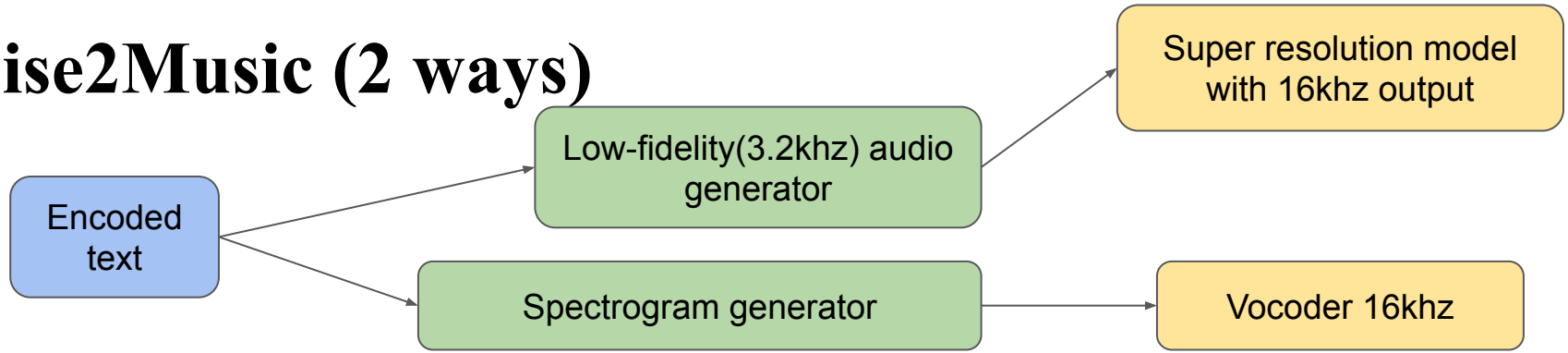
1. R - 1D ResNet
2. M - modulus + noise level conditioning
3. I - inject latent representation
4. A - self-attention
5. C - cross attention with text embedding

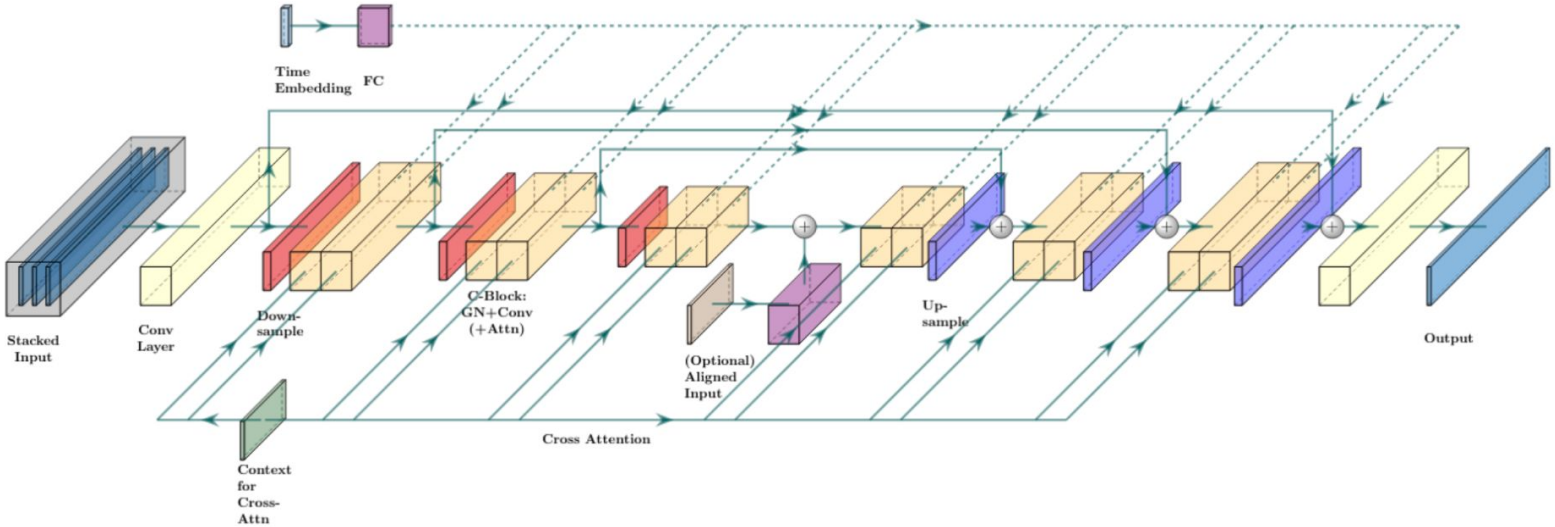
Mousai: Efficient Text-to-Music Diffusion





Noise2Music (2 ways)





Сравнение по метрикам

Table 1: Text-to-Music generation. We compare objective and subjective metrics for MUSICGEN against a number of baselines. We report both mean and CI95 scores. The Mousai model is retrained on the same dataset, while for MusicLM we use the public API for human studies. We report the original FAD on MusicCaps for Noise2Music and MusicLM. "MUSICGEN w. random melody" refers to MUSICGEN trained with chromagram and text. At evaluation time, we sample the chromagrams at random from a held-out set.

	MUSICCAPS Test Set				
MODEL	FAD$_{vgg}$ ↓	KL ↓	CLAP$_{scr}$ ↑	OVL. ↑	REL. ↑
Riffusion	14.8	2.06	0.19	79.31±1.37	74.20±2.17
Mousai	7.5	1.59	0.23	76.11±1.56	77.35±1.72
MusicLM	4.0	-	-	80.51±1.07	82.35±1.36
Noise2Music	**2.1**	-	-	-	-
MUSICGEN w.o melody (300M)	3.1	1.28	0.31	78.43±1.30	81.11±1.31
MUSICGEN w.o melody (1.5B)	3.4	1.23	**0.32**	80.74±1.17	**83.70**±1.21
MUSICGEN w.o melody (3.3B)	3.8	**1.22**	0.31	**84.81**±0.95	82.47±1.25
MUSICGEN w. random melody (1.5B)	5.0	1.31	0.28	81.30±1.29	81.98±1.79

Выводы

- MusicGen использует трансформер с несколькими потоками сжатого дискретного музыкального представления, в то время как бейзлайны используют либо каскадные или иерархические модели
- MusicGen не требует дополнительной предобученной семантической репрезентации, в то время как бейзлайны используют либо предобученные языковые модели, либо специальные эмбеддинги для текста или мелодии
- MusicGen может генерировать как моно, так и стерео образцы, в то время как бейзлайны ограничены только моно форматом
- MusicGen позволяет лучше контролировать генерируемый вывод, используя условия на текстовое описание или мелодические особенности, в то время как бейзлайны либо не поддерживают условную генерацию, либо имеют только текст

References

- https://arxiv.org/pdf/2306.05284.pdf - MusicGen
- https://www.riffusion.com - Riffusion
- https://arxiv.org/pdf/2301.11325.pdf - MusicLM
- https://arxiv.org/pdf/2301.11757.pdf - Mousai
- https://arxiv.org/pdf/2302.03917.pdf - Noise2Music

Вопросы?