Generating Symbolic Music From Natural Language Prompts Using an LLM-Enhanced Dataset

Weihan Xu[1] **Julian McAuley**[2] **Taylor Berg-Kirkpatrick**[2] **Shlomo Dubnov**[2] **Hao-Wen Dong**[3]

[1]Duke University [2]University of California San Diego [3]University of Michigan

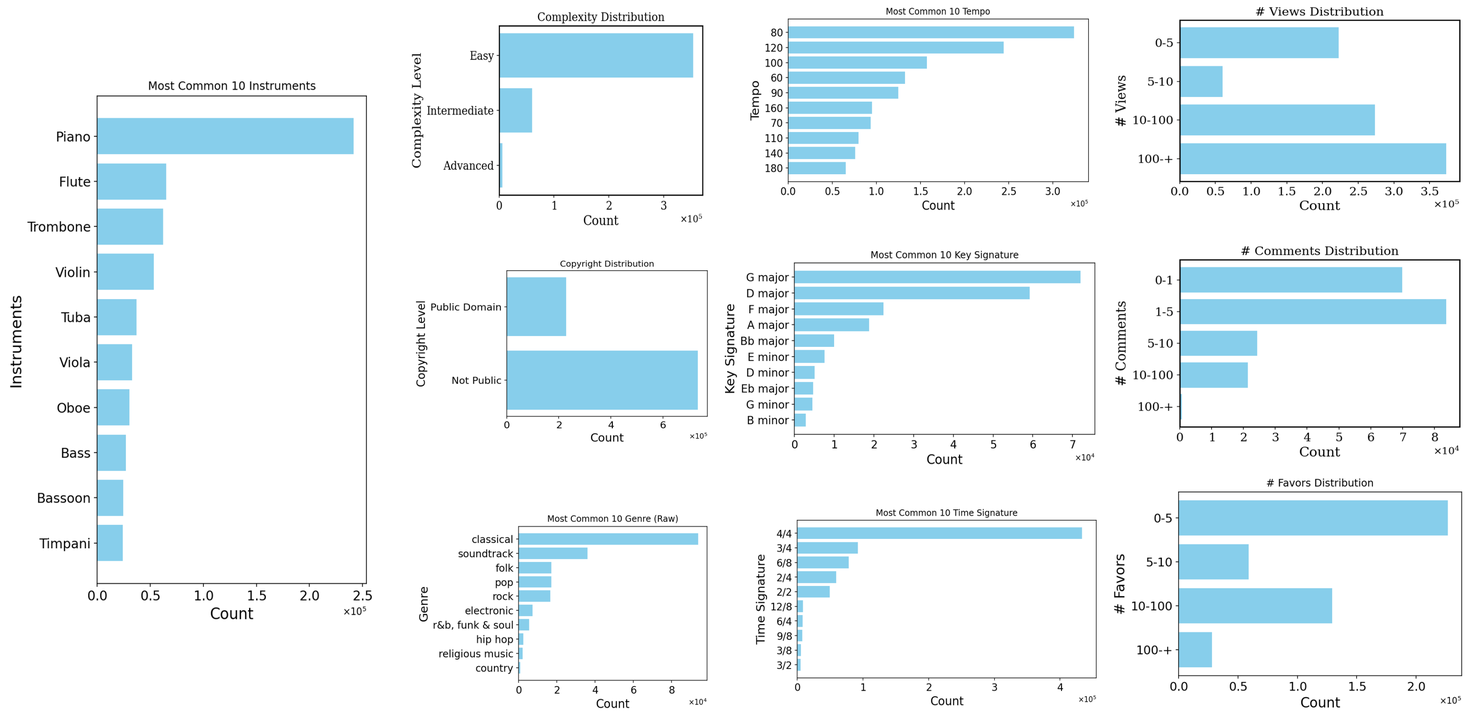
Overview

Symbolic Music Generation systems generate music in editable formats that can be further completed by the users, making it easier for musicians to integrate such systems to creative workflow. However, symbolic-domain controllable music generation has lagged behind partly due to the lack of a large-scale symbolic music dataset with extensive metadata and captions

Contribution

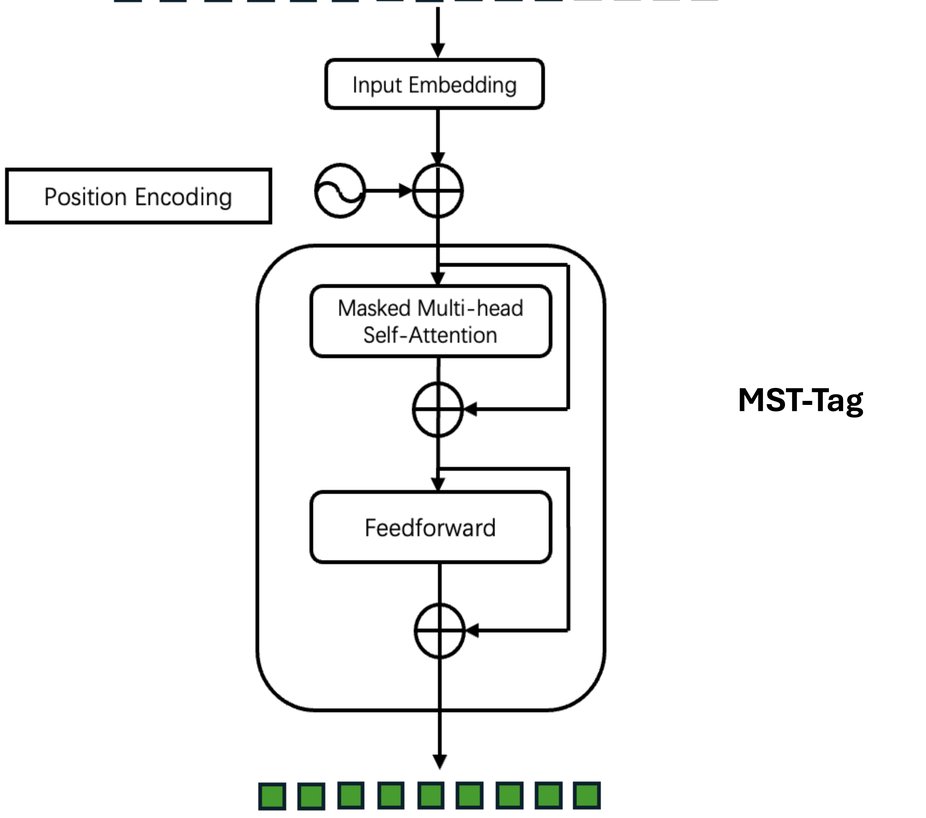
- We propose **MetaScore**, a new publicly available dataset with musical scores paired with **rich metadata** and **LLM-generated natural language captions**.
- We train two new models for **tag-** and **text-based controllable symbolic music generation** that support **instrument, genre, composer and complexity** controls.

MetaScore

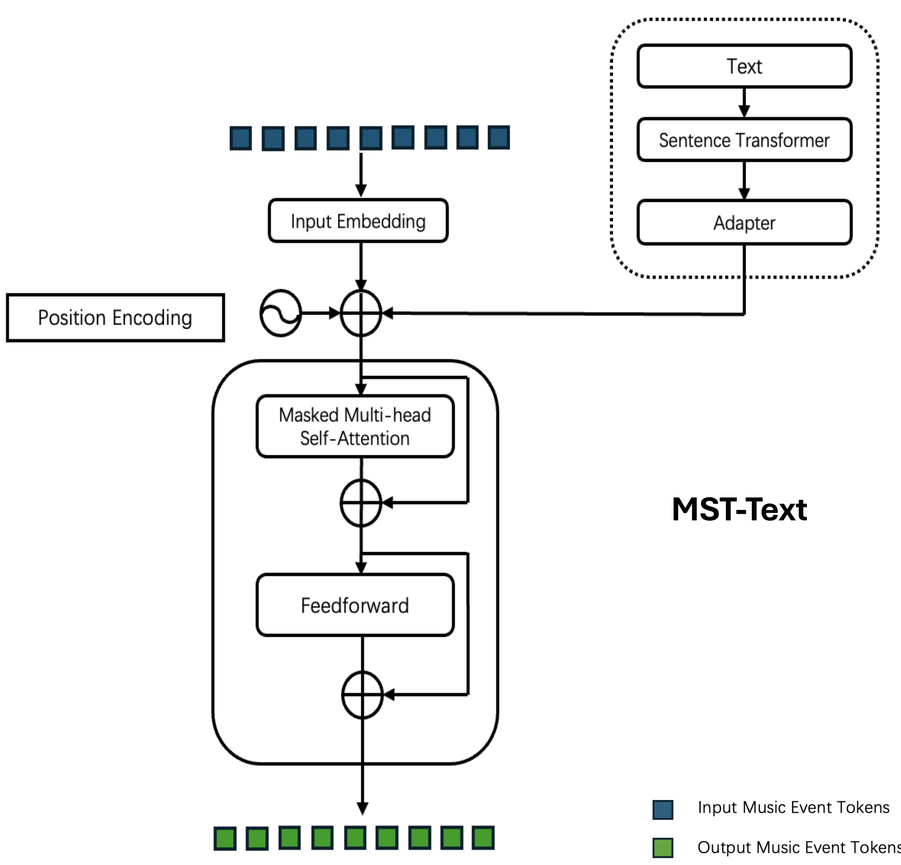


- **963K** raw MuseScore file and MusicXML
- **Sources**: MuseScore Forum
- **Annotations:** Genre, Composer, Complexity, Key signature, Time Signature, Tempo, User Interaction Statistics.

MetaScore Transformer



MST-Tag



MST-Text

Results

Objective Evaluation Results

	Pitch class entropy	Scale consistency	Groove consistency
MST-Tags-Small	2.88 ± 0.08	0.89 ± 0.02	**0.92 ± 0.01**
MST-Tags	2.93 ± 0.07	0.89 ± 0.02	0.90 ± 0.01
BART-based [13]	2.54 ± 0.06	0.99 ± 0.00	1.00 ± 0.00
MST-Text	**2.70 ± 0.06**	**0.95 ± 0.01**	**0.92 ± 0.01**
Ground truth	2.67 ± 0.06	0.95 ± 0.01	0.92 ± 0.01

Objective evaluation results on music quality with conditions from MST test set

Subjective Evaluation Results

	Model size	Training samples	Coherence↑	Arrangement↑	Adherence↑	Overall quality↑
MST-Tags-Small	87.36M	150K	3.87 ± 0.36	3.98 ± 0.38	**3.86 ± 0.38**	3.57 ± 0.37
MST-Tags	87.36M	901K	**4.01 ± 0.37**	**4.06 ± 0.39**	3.60 ± 0.49	3.66 ± 0.45
BART-based [13]	139M	283K	3.86 ± 0.30	3.63 ± 0.39	2.81 ± 0.50	3.29 ± 0.42
MST-Text	87.44M	560K	3.93 ± 0.28	3.88 ± 0.33	3.35 ± 0.44	**3.69 ± 0.33**

Subjective Evaluation Results in terms of Coherence, Arrangement, Adherence and Overall Quality in a Likert Scale of 1 to 5

Model	size	CLAP Score↑ M	CLAP Score↑ T	CLAP Score↑ M+T	Coherence(%)↑ M	Coherence(%)↑ T	Coherence(%)↑ M+T	Arrangement(%)↑ M	Arrangement(%)↑ T	Arrangement(%)↑ M+T	Adherence(%)↑ M	Adherence(%)↑ T	Adherence(%)↑ M+T	Overall quality(%)↑ M	Overall quality(%)↑ T	Overall quality(%)↑ M+T
Text2MIDI [1]	159M	0.23	0.20	0.22	0	40	20	40	**50**	45	40	**80**	**60**	40	**60**	**50**
MST-Text	87.44M	**0.36**	0.13	**0.24**	**100**	**60**	**80**	**60**	**50**	**55**	**60**	20	40	**60**	40	**50**

Comparison of MST-Text and Text2MIDI on three prompt sets: 1) M: five prompts from our test set; 2) **T**: five prompts from the Text2MIDI; 3) **M+T**: the union of these two prompt sets

Paper

Demo

Project page

https://wx83.github.io/MetaScore_Official/

Paper

https://arxiv.org/pdf/2410.02084