A NEW DATASET FOR TAG- AND TEXT-BASED CONTROLLABLE SYMBOLIC MUSIC GENERATION

Weihan Xu[1] **Julian McAuley**[2] **Taylor Berg-Kirkpatrick**[2]
Shlomo Dubnov[2] **Hao-Wen Dong**[2,3]
[1] Duke University [2] University of California San Diego [3] University of Michigan
weihan.xu@duke.edu

ABSTRACT

Recent years have seen many audio-domain text-to-music generation models that rely on large amounts of text-audio pairs for training. However, similar attempts for symbolic-domain controllable music generation has been hindered due to the lack of a large-scale symbolic music dataset with extensive metadata and captions. In this paper, we introduce *MetaScore*, a novel dataset of 963K musical scores, along with extensive metadata collected from an online music forum. Additionally, we provide machine-generated captions for each score. With MetaScore, we explore controllable symbolic music generation and showcase the potential of our proposed dataset in enabling generating symbolic music using free-form natural language.

1 MetaScore Dataset

1.1 Collecting and Preprocessing the Dataset

We collected 962,586 music scores from MuseScore, each paired with its corresponding metadata. We refer to this original dataset as *MetaScore-Raw*. MetaScore-Raw includes extensive metadata such as genre, composer, complexity, key signature, time signature, tempo, and user interaction statistics (e.g., number of views, likes, and comments). From the raw MSCZ files, we extract musical instruments, retaining only those compatible with the General MIDI standard. We also extract the time signature, key signature, and tempo from MusicXML files.

1.2 Inferring Missing Genre Tags in MetaScore-Raw

While MetaScore-Raw provides rich metadata information, we notice that not all songs come with complete metadata. For example, only 181K (18.8%) out of 963K songs in MetaScore-Raw contain genre metadata. As genre is one of the most intuitive ways for a user to control the style for a music generation system, we want to complete the genre information for songs without a genre label in MetaScore-Raw.

The genre tagger is based on the Multitrack Music Transformer (MMT) [1], where we remove the casual mask used for autoregressive modeling and append a multi-label classification layer. We select the threshold of the multi-label classification layer for each class based on the F1 score on the validation set. To evaluate the performance of genre tagging, we first compute the precision, recall and F1-score on the test set, where we achieve a micro-averaged precision of 61.94, recall of 63.03, and F1 score of 62.48. In addition, we conducted a subjective listening test to compare the quality of the auto-generated genre tags with the user-annotated tags in MetaScore-Raw. The 22 participants are instructed to answer the following question in a Likert scale of 1 to 5: *"How well do you think this piece of music aligns with the following genre?"*. From Table 1, we can see that the auto-generated genre tags in MetaScore-Plus achieves a lower tags-music adherence compared to the ground truth tags in MetaScore-Genre, but the difference did not reach statistical significance in our setup.

Type	Dataset	Adherence↑
Ground truth genre tags	MetaScore-Genre	3.11 ± 0.49
Auto-generated genre tags	MetaScore-Plus*	3.05 ± 0.54
LLM-generated captions	MetaScore-Plus	3.23 ± 0.49

* We only include songs with auto-generated genre tags here.

Table 1. Subjective evaluation results on tags/text-music adherence of the dataset in a Likert scale of 1 to 5. We report the mean values and 95% confidence intervals.

1.3 Generating Pseudo Captions using LLMs

To enable text-based downstream tasks (e.g., music captioning and text-to-music generation), we leverage large language models to convert the metadata into natural language captions. We follow LP-MusicCaps [2] and CLAP [3] and adopt an in-context learning-based approach [4].

We form the input prompt string by combining genre, composer, time signature, key signature, tempo, complexity, and user-specified tags.

We provide six examples followed by the input tags that need to be inferred, and the large language model (specifically, Bloom [5]) is expected to generate the pseudo cap-

Late-Breaking / Demo Session Extended Abstract, ISMIR 2024 Conference

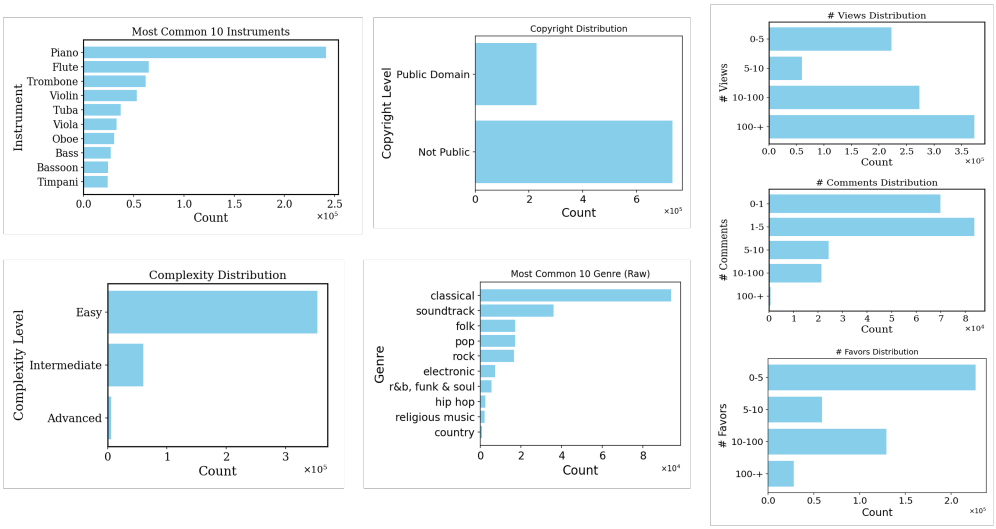

Figure 1. Statistics of the metadata available in MetaScore. Note that not all songs come with complete metadata.

	Dataset	Input type	Model size	Training samples	Coherence↑	Arrangement↑	Adherence↑	Overall quality↑
MST-Tags-Small	MetaScore-Genre	Tag	87.36M	150K	3.87 ± 0.36	3.98 ± 0.38	$\mathbf{3.86 \pm 0.38}$	3.57 ± 0.37
MST-Tags	MetaScore-Plus	Tag	87.36M	901K	$\mathbf{4.01 \pm 0.37}$	$\mathbf{4.06 \pm 0.39}$	3.60 ± 0.49	3.66 ± 0.45
MST-Text	MetaScore-Plus	Text	87.44M	560K	3.93 ± 0.28	3.88 ± 0.33	3.35 ± 0.44	$\mathbf{3.69 \pm 0.33}$

Table 2. Subjective evaluation results in a Likert scale of 1 to 5. We report the mean values and 95% confidence intervals.

tions based on these input tags. We generate the pseudo captions using the Hugging Face API [6]. We truncate the output sequence to a maximum of 32 tokens.

To evaluate the quality of the generated pseudo captions, we also include this dataset in Table 1 conduct a subjective listening test, to evaluate whether the quality of the auto-generated genre tags match given text descriptions in a Likert scale of 1 to 5. As shown in Table 1, the LLM-generated captions achieve the highest tags/text-music adherence across datasets, possibly because the participants prefer a natural language caption than a list of tags.

1.4 Versions of MetaScore

We will release the following five versions of MetaScore:

- *MetaScore-Raw* (963K): The raw MuseScore files and metadata scraped from the MuseScore forum.
- *Metascore-Genre* (181K): A subset of MuseScore-Raw containing files with user-annotated genres. Additionally, we discard any songs composed by a composer that has less than 100 compositions in MetaScore-Raw. We also provide LLM-generated captions based on information extracted from the metadata in Metascore-Genre.
- *MetaScore-Plus* (963K): MetaScore-Raw where missing genre tags are completed by the trained genre tagger described in Section 1.2. We also provide LLM-generated captions based on information extracted from the metadata in MetaScore-Plus.

We will release all music scores and metadata that are in the public domain or with a Creative Commons licenses. The rest of the dataset will be available upon request for research purpose.

2 Discussion and Conclusion

In this work, we introduce MetaScore, a new large symbolic dataset with rich metadata and pseudo captions, offering two versions: MetaScore-Genre and MetaScore-Plus. MetaScore-Genre excludes uncommon composers and auto-generated tags, aiming for cleaner metadata, while MetaScore-Plus provides a more comprehensive dataset.

With MetaScore, we conduct text-conditioned music generation and tag-conditioned music generation. We evaluate the impact of these two versions on tag-and text-conditioned music generation using models MST-Tags-Small (trained on MetaScore-Genre) and MST-Tags (trained on MetaScore-Plus), along with MST-Text, a text-conditioned model. Notice that the LLM-generated captions that we evaluate in Table 1 and the training of MST-text are using an earlier version of MetaScore. In that early version of MetaScore, key signature, time signature and tempo are not considered. Please refer to our demo page for listening samples. [1]

Table 2 indicates that MST-Tags excels in coherence and arrangement despite its reduced adherence, likely due to the inclusion of auto-generated tags. Meanwhile, MST-Tags-Small scores highest in adherence, benefiting from its cleaner dataset. MST-Text stands out in overall quality, possibly because of the natural language-based user interface it provides.

[1] https://geniusmusic.github.io/ISMIR2024/

3 References

[1] H.-W. Dong, K. Chen, S. Dubnov, J. McAuley, and T. Berg-Kirkpatrick, "Multitrack music transformer," in *Proc. ICASSP*, 2023.

[2] S. Doh, K. Choi, J. Lee, and J. Nam, "LP-MusicCaps: LLM-based pseudo music captioning," in *Proc. ISMIR*, 2023.

[3] Y. Wu, K. Chen, T. Zhang, Y. Hui, M. Nezhurina, T. Berg-Kirkpatrick, and S. Dubnov, "Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation," in *Proc. ICASSP*, 2023.

[4] Q. Dong, L. Li, D. Dai, C. Zheng, Z. Wu, B. Chang, X. Sun, J. Xu, L. Li, and Z. Sui, "A survey on in-context learning," in *arXiv preprint arXiv:2301.00234*, 2023.

[5] T. L. Scao, A. Fan, C. Akiki, E. Pavlick, S. Ilić, D. Hesslow, and et al., "Bloom: A 176b-parameter open-access multiliguai language model," in *arXiv preprint arXiv:2211.05100*, 2022.

[6] Hugging Face, "Hugging face api," https://huggingface.co/bigscience/bloom, 2024, accessed: 2024-08-19.