# Supplementary Material for: ”Performance Conditioning for Diffusion-Based Multi-Instrument Music Synthesis”

IEEE Publication Technology, *Staff, IEEE,*

## I. Classifier-Free Guidance with Multiple Conditions

In our case of performance-aware multi-instrument synthesis, we apply two simultaneous conditions: A note condition defining the notes to be played, and a performance condition defining the acoustic properties such as timbre, recording environment, etc. It is possible to apply CFG using multiple conditions in a straightforward manner: Denote by $c_1, c_2$ the two types of conditions. During training, we dropout each of the conditions $c_1, c_2$ with probability $0.1$, independently. Then, during sampling, we define:

$$\begin{aligned}
\epsilon_{cond1} &= \epsilon_\theta(x_t, t, c1, 0) \\
\epsilon_{cond2} &= \epsilon_\theta(x_t, t, 0, c2) \\
\epsilon_{cond1,2} &= \epsilon_\theta(x_t, t, c1, c2) \qquad (1) \\
\epsilon &= \epsilon_{cond12} + (1 - w_1)(\epsilon_{cond1,2} - \epsilon_{cond2}) \\
&\quad + (1 - w_2)(\epsilon_{cond1,2} - \epsilon_{cond1})
\end{aligned}$$

where $w_1, w_2 > 1$ are extrapolation weights representing the strengths of the two conditions, respectively. Intuitively, the term $(1 - w_1)(\epsilon_{cond1,2} - \epsilon_{cond2})$ enhances the condition $c_1$ while remaining faithful to the condition $c_2$, and the term $(1 - w_2)(\epsilon_{cond1,2} - \epsilon_{cond1})$ enhances the condition $c_2$ while remaining faithful to the condition $c_1$. We note that other ways to perform the extrapolation are possible, such as using $\epsilon_\theta(x_t, t, 0, 0)$ (zeroing out both conditions), but the approach described in Equation 1 gave best results in practice.

## II. Dataset Distribution

The instrumentation distribution of the datasets used for training and evaluation can be seen in Tables I (train), II (listening test), and III (large evaluation set).

## III. Quantitative Results over Large Scale Evaluation Set

In this section, we compare different models trained on the same dataset (our 58-hour dataset, Table I), and perform a quantitative evaluation on larger scale dataset, of over 5 hours (Table III).

We compare the following models: U-Net and T5 models trained with alignments as note conditions, and a T5 model trained with transcriptions as note conditions. For the latter model we provided two forms of note conditions: Pitch with instrument, and pitch-only. The transcriptions were obtained by training a transcriber [1, 2] on the pairs of audio and alignments, and providing the (thresholded) predictions as note conditions to the synthesizer, rather than the alignments.

TABLE I
**Train data** distribution. The train dataset contains audio and MIDI. We show the total length for each ensemble (Length), and how many performances in the train set are played by each ensemble (#Performances).

| Ensemble | Length | #Performances |
|---|---|---|
| Flute & Harpsichord | 0:49:19 | 2 |
| Orchestra | 14:31:13 | 39 |
| Orchestra & Choir | 1:49:54 | 1 |
| Orchestra & Piano | 5:12:50 | 7 |
| Solo Cello | 2:16:16 | 2 |
| Solo Flute | 0:09:16 | 1 |
| Solo Guitar | 4:40:14 | 85 |
| Solo Harpsichord | 7:27:40 | 12 |
| Solo Organ | 2:10:49 | 25 |
| Solo Piano | 5:52:41 | 17 |
| Solo Violin | 1:56:49 | 1 |
| Violin & Harpsichord | 3:18:48 | 2 |
| Violin & Piano | 3:59:17 | 1 |
| Violin, Cello, & Piano | 3:37:23 | 1 |
| Wind Quintet | 0:33:10 | 1 |
| **All** | 58:25:47 | 197 |

Together, 4 different configurations. We train all models both with, and without performance conditioning – together 8 configurations.

Results are reported in Tables IV (Group-FAD), V (performance classification accuracy), VI (All-FAD), and VII (transcription accuracy). The general trend we clearly and consistently observe is that Group-FAD and classification accuracy dramatically improve as a result of performance conditioning, while All-FAD and transcription metrics remain comparable (possibly with a slight increase or decrease). That is, by incorporating performance conditioning, we can generate performances with the same notes, but such performances that more resemble the desired target performances, and do so while maintaining quality. We describe the results in detail:

*a) All-FAD (Large-Scale):* One can see in Table VI that performance conditioning does not significantly impact the All-FAD – the VGGish All-FAD slightly increases while the TRILL All-FAD slightly decreases. Increase in All-FAD as a result of conditioning is not necessarily surprising – conditioning forces the generated performances to deviate from the general distribution and more resemble specific performances

This paper was produced by the IEEE Publication Technology Group. They are in Piscataway, NJ.

Manuscript received April 19, 2021; revised August 16, 2021.

TABLE II
MIDI performances used for the listening test. The number in parentheses represents the movement. The first 3 minutes of each performance were used.

| MIDI | Ensemble |
|---|---|
| Bach Badineri | wind quintet |
| Bach Italian Concerto (movement 1) | harpsichord |
| Bach Toccata and Fugue | church organ |
| Bach Orchestral Suite 1 (1) | orchestra |
| Mozart Symphony 40 (movement 1) | orchestra |
| Beethoven Symphony 5 (movement 1) | orchestra |
| Beethoven Symphony 6 (movement 1) | orchestra |
| Beethoven Symphony 6 (movement 3) | orchestra |
| Bach Mass in B Minor (movement 11) | choir & orchestra |
| Mozart Piano Concerto 20 (movement 1) | piano & orchestra |

TABLE III
**Test data** distribution. The test dataset contains MIDI only, without corresponding audio. We show the total length for each ensemble (Length), and how many MIDI performances in the test set are played by each ensemble (#Performances.). As described in the paper, each test MIDI is synthesized 3 times, with 3 different random performance conditions from the train set, from its corresponding ensemble.

| Ensemble | Length | #Performances |
|---|---|---|
| Flute & Harpsichord | 0:04:11 | 1 |
| Orchestra | 2:16:02 | 19 |
| Orchestra & Choir | 0:11:49 | 4 |
| Orchestra & Piano | 0:52:27 | 3 |
| Solo Cello | 0:05:53 | 3 |
| Solo Guitar | 0:05:07 | 3 |
| Solo Harpsichord | 0:09:17 | 4 |
| Solo Organ | 0:07:33 | 1 |
| Solo Piano | 0:18:53 | 2 |
| Solo Violin | 0:13:14 | 5 |
| Violin & Harpsichord | 0:05:44 | 2 |
| Violin, Cello, & Piano | 0:34:29 | 9 |
| Wind Quintet | 0:04:51 | 2 |
| **All** | 5:09:30 | 58 |

that appear in the conditions. However, All-FAD measures similarity in distribution to the entire train set as a whole, which relates more to general quality and realism.

*b) Group-FAD, Classification (Large-Scale):* The dramatic effect of performance conditioning can be seen in Tables IV and V. The Group-FAD metric dramatically improves as a result of performance conditioning, for all models. This means the distribution of the generated performances becomes perceptually more similar to the conditioning performances.

Next, we observe the classification accuracy (Table V). We classify each generated performance to the performance in the train set that is its Group-FAD nearest, and measure for how many of the generated performances this yields the target conditioning performance. It can be seen that performance conditioning improves classification accuracy by over 30%, reaching near 90% top-3 accuracy (out of 197 reference performances). For example, as shown in Table V, without performance conditioning the top-1 classification accuracy is only 35.5% for the model T5 (Aligned). When using performance conditioning, it increases dramatically to 68%. Similar improvements can also observed for the U-Net, and when looking at top-3 accuracy values. Similarly to the Group-FAD results, these results suggest that performance conditioning helps adapt to the specific timbre and room acoustics of a performance.

*c) Transcription (Large-Scale):* We observe the transcription metrics (Table VII), measuring if the synthesized performances actually realize the notes specified by the MIDI note condition. There is no entirely objective or absolute way to measure this, however, we can still gain insights by using an automatic transcriber. We therefore use a transcriber trained on precisely the same data as the synthesizer. Note that such metrics are influenced not only by the quality of the synthesizer, but also from the quality of the transcriber.

In Table VII we can observe that most models yield similar transcription metrics, whether using performance conditioning or not, reaching accuracy of up to 67% (note-level), which is of reasonable magnitude when considering the complexity of highly polyphonic orchestral music. In addition, we can observe that using transcriptions as conditions rather than alignments provides better overall transcription metrics.

It can also be seen that unsurprisingly, the note-with-instrument metric is significantly lower when using pitch-only input. Note however that this is significantly mitigated when using performance conditioning (15% improvement). This indicates that performance conditioning helps achieving the target instrumentation, even when using pitch-only note conditions, as it learns the correlations between performances and their instrumentation.

### A. *Difference in All-FAD vs. Group-FAD Value Range*

When comparing the actual FAD values in Tables VI, IV, one can see the All-FAD values are lower than the Group-FAD. We attribute this to the fact that the mean vectors and covariance matrices for All-FAD are computed over significantly larger evaluation and reference datasets than for Group-FAD and therefore yield less statistical fluctuations, resulting in an overall lower FAD score.

### B. *Supplementary Project Page*

While the quantitative analysis in this section indicates that performance conditioning indeed improves perceptual similarity in music synthesis, it cannot replace listening to the actual generated samples. Therefore, in order to assess the potential of performance conditioning in the context of diffusion-based music synthesis, we strongly encourage the reader to

TABLE IV
Group-FAD results on large evaluation set. Best Result Is Bold, and Next-Best Underlined. In all models, performance conditioning dramatically improves the Group-FAD, implying the generated performances more resemble the targets. Results are comparable for all models (except for pitch-only conditions), with alignment producing slightly better scores. Note the dramatic improvement achieved by performance conditioning when using pitch-only input.

| | Group-FAD↓ | | | |
|---|---|---|---|---|
| | VGGish | | TRILL | |
| P Con. | w/o | w/ | w/o | w/ |
| U-Net Aligned | 7.06 | 5.21 | 0.5 | 0.33 |
| T5 Aligned | 6.95 | 5.46 | 0.51 | 0.35 |
| T5 Transcribed | 7.46 | 5.68 | 0.55 | 0.36 |
| T5 Transcribed Pitch | 12.81 | 6.03 | 0.7 | 0.38 |

TABLE V
Performance classification results on large evaluation set. Results are consistent with Table IV.

| | Classification% Top-1/3/5↑ | |
|---|---|---|
| P Con. | w/o | w/ |
| U-Net Aligned | 11.6/24.5/36.1 | 52.9/73.5/84.5 |
| T5 Aligned | 35.5/60.0/69.7 | 67.7/89.7/91.0 |
| T5 Transcribed | 16.8/31.6/41.9 | 66.5/83.9/88.4 |
| T5 Transcribed Pitch | 4.5/18.1/27.1 | 62.6/79.4/85.8 |

TABLE VI
All-FAD results on large evaluation set. Results are comparable for all models (except for pitch-only conditions), with alignment producing slightly better scores. Note that in most cases performance conditioning does not significantly impact All-FAD scores, since this metric measures general similarity to the train set. However, it does improve the scores for pitch-only input.

| | All-FAD↓ | | | |
|---|---|---|---|---|
| | VGGish | | TRILL | |
| P Con. | w/o | w/ | w/o | w/ |
| U-Net Aligned | 3.37 | 3.94 | 0.12 | 0.11 |
| T5 Aligned | 3.98 | 3.53 | 0.12 | 0.09 |
| T5 Transcribed | 3.05 | 3.58 | 0.12 | 0.11 |
| T5 Transcribed Pitch | 3.97 | 3.78 | 0.18 | 0.12 |

TABLE VII
Transcription results on large evaluation set. Results are comparable for all models, except for note-with-instrument for pitch-only input. Performance conditioning significantly improves this metric for pitch-only input, implying that performance conditioning can somewhat replace instrument conditioning. Note that transcription produces slightly better scores than alignment.

| | Transcription F1% ↑ | | | | | |
|---|---|---|---|---|---|---|
| | Note | | Note & Inst. | | Frame | |
| P Con. | w/o | w/ | w/o | w/ | w/o | w/ |
| U-Net Aligned | 63.1 | 61.8 | 46.6 | 46.2 | 65.4 | 64.1 |
| T5 Aligned | 62.0 | 63.2 | 38.4 | 47.1 | 60.7 | 62.0 |
| T5 Transcribed | 66.9 | 64.7 | 50.7 | 46.2 | 64.8 | 63.9 |
| T5 Transcribed Pitch | 67.2 | 64.3 | 25.0 | 40.3 | 64.7 | 63.2 |

listen to the samples provided on our supplemental website (benadar293.github.io/midipm). We provide comparisons of MIDI files sonified with a simple concatenative synthesizer, the baseline approach [3], and our proposed method. Furthermore, we show the conditioning effect when synthesizing the same MIDI file with a variety of different performance conditions. Among others, we render Bach's 8th Invention on eight different harpsichords, and Beethoven's Pastoral Symphony with four different orchestras and recording environments.

## IV. Discussion

We presented a framework for training neural synthesizers on real performances, using diffusion models conditioned on notes and a performing style. We demonstrated that the latter condition both improves realism of multi-instrument performances of classical music, including orchestral symphonies, and adapts to the specific characteristics of a given performance, such as timbre and recording environment. Important future work includes extension to other genres, such as jazz, ethnic, pop music, and even human singing. Another important direction is exploring other spectral domains, such as STFT, CQT, etc. Yet another direction involves human speech – we believe a unified diffusion-based framework for music and speech is possible, by providing additional textual or phonemic conditions.

## References

[1] C. Hawthorne, A. Stasyuk, A. Roberts, I. Simon, C. A. Huang, S. Dieleman, E. Elsen, J. H. Engel, and D. Eck, "Enabling factorized piano music modeling and generation with the MAESTRO dataset," in *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. [Online]. Available: https://openreview.net/forum?id=r1lYRjC9F7

[2] B. Maman and A. H. Bermano, "Unaligned supervision for automatic music transcription in the wild," in *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, ser. Proceedings of Machine Learning Research, K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvári, G. Niu, and S. Sabato, Eds., vol. 162. PMLR, 2022, pp. 14 918–14 934. [Online]. Available: https://proceedings.mlr.press/v162/maman22a.html

[3] C. Hawthorne, I. Simon, A. Roberts, N. Zeghidour, J. Gardner, E. Manilow, and J. H. Engel, "Multi-instrument music synthesis with spectrogram diffusion," in *Proceedings of the 23rd International Society for Music Information Retrieval Conference, ISMIR 2022, Bengaluru, India, December 4-8, 2022*, 2022, pp. 598–607. [Online]. Available: https://archives.ismir.net/ismir2022/paper/000072.pdf