# Contin-U: Full-Score to Performance Audio with Cross-Attentive System-Continuation Inference

Jongmin Jung
*Dept. of Artificial Intelligence*
*Sogang University*
Seoul, Republic of Korea
jongmin@sogang.ac.kr

Dongmin Kim
*Dept. of Artificial Intelligence*
*Sogang University*
Seoul, Republic of Korea
dmkim@sogang.ac.kr

Sihun Lee
*Dept. of Artificial Intelligence*
*Sogang University*
Seoul, Republic of Korea
sihunlee@sogang.ac.kr

Seola Cho
*Dept. of Art & Technology*
*Sogang University*
Seoul, Republic of Korea
seola.cho@sogang.ac.kr

Hyungjoon Soh
*Department of Physics Education*
*Seoul National University*
Seoul, Republic of Korea
hjsoh88@snu.ac.kr

Irmak Bukey
*Computer Science Department*
*Carnegie Mellon University*
Pittsburgh, PA, USA
ibukey@andrew.cmu.edu

Chris Donahue
*Computer Science Department*
*Carnegie Mellon University*
Pittsburgh, PA, USA
chrisdonahue@cmu.edu

Dasaem Jeong
*Dept. of Art & Technology*
*Sogang University*
Seoul, Republic of Korea
dasaemj@sogang.ac.kr

***Abstract*—We present a system for rendering full-length, expressive performance audio from MusicXML scores by adapting a unified cross-modal Transformer. Our approach circumvents direct symbolic processing; instead, we engrave the input score to images and leverage the model's direct image-to-audio synthesis pathway, which generates expressive audio without requiring an intermediate symbolic performance representation. To scale from short training crops (1–3 systems) to complete works, we introduce a lightweight inference-time method: a two-system sliding window that uses cross-attention analysis to detect precise system boundaries in the audio output. This allows the fixed model backbone to produce seamless, long-form audio without any retraining.**

## I. Introduction

Expressively translating written music into audio remains a core MIR goal. Our earlier work *U-MusT: Unified Cross-modal Translation of Score Images, Symbolic Music, and Performance Audio* [1] unifies score images, symbolic notation, MIDI, and performance audio within a single encoder–decoder Transformer trained on a 1,300-hour paired corpus with additional symbolic resources, achieving both the first score-image-conditioned audio generation and a marked reduction in OMR error.

Running the model in its image-to-audio direction yields implicit performance modelling: phrasing, rubato, and dynamics emerge directly from engraved images, with no explicit symbolic performance layer. We exploit this property for RenCon 2025 by engraving the provided MusicXML scores into page images and feeding them to the unchanged image-to-audio branch.

A simple inference-time sliding window, described next, then scales the fixed backbone from short training crops to complete pieces.

## II. Methodology

We employ the unified sequence-to-sequence Transformer from *U-MusT*, trained jointly on three directions: Optical Music Recognition, MIDI-to-Audio, and image-to-audio. All modalities are serialized as token sequences: (i) images and audio via residual vector quantization (RVQ) codebooks (RQ-VAE [2] for images, DAC [3] for audio), and (ii) symbolic MusicXML and MIDI via linearized MusicXML(LMX) [4] and MIDI-like token [5], [6] vocabularies. During training, multiple consecutive system image crops can be concatenated with a special `<SEP>` token.

### *A. Preprocessing from MusicXML to Image Tokens*

Given that the RenCon task supplies MusicXML while our inference path consumes system image tokens, we convert as follows (no additional stages retained):

1) Render each input MusicXML score to high-resolution engraved page images using MuseScore [7].
2) Detect musical system regions on each rendered page using our fine-tuned YOLOv8-medium [8] detector (`YOLO-system`); crop each system.
3) Estimate staff height for each system crop with a second YOLOv8-medium model (`YOLO-staff`); resize the crop so its staff-height distribution matches that of the translation model's training data.
4) Tokenize each normalized system crop with the trained RQVAE, producing discrete RVQ image token sequences. Consecutive system token sequences are concatenated with a single `<SEP>` token between them when forming two-system windows for inference.

### *B. System-Continuation Inference*

The model was trained only on short windows (1–3 systems). To generate full-length performances we slide a two-system window $(S_j, S_{j+1})$ across the score.

For each window we:

- Form the input token sequence by concatenating the RVQ tokens of $S_j$, an explicit `<SEP>`, then those of $S_{j+1}$, along with positional and modality indices.

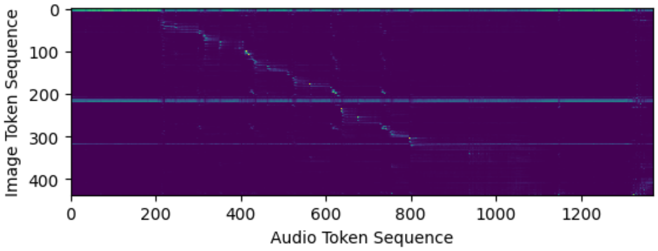

Fig. 1. Cross-attention matrix $A$ (x: audio tokens, y: flattened image tokens). The horizontal band at row $i_{\text{sep}} = 211$ marks the `[SEP]` token; diagonal energy traces indicate monotonic image–audio alignment up to the boundary before attention shifts past `[SEP]`.

- Run autoregressive image-to-audio generation. For the first window ($j = 0$) no previously generated audio prefix is supplied. For every subsequent window ($j > 0$) we supply, as conditioning prefix, the audio token segment previously generated for the second system of the preceding window (i.e., the portion aligned with $S_j$ in window $(S_{j-1}, S_j)$). During each forward pass we capture cross-attention tensors from selected heads in the last decoder layer, which exhibit distinct near-diagonal alignment structure.
- Let the cross-attention matrix be $A \in \mathbb{R}^{I \times T}$ with rows (flattened image tokens) $i = 1, \ldots, I$ and columns (audio token steps) $t = 1, \ldots, T$, and let $i_{\text{sep}}$ denote the [SEP] image token row. For each audio step $t$, define $p_t = \sum_{i > i_{\text{sep}}} A_{i,t}$, the proportion of attention mass on image tokens after the separator. The first $t$ for which $p_t$ exceeds a threshold (e.g. $0.5$) for at least three consecutive steps is taken as the boundary separating the audio aligned with $S_j$ from that aligned with $S_{j+1}$.
- Slice and accumulate the audio tokens up to the boundary into the global sequence and use the following run of tokens as the conditioning prefix for the next window.
- Repeat for the next overlapping window $(S_{j+1}, S_{j+2})$ until all systems are processed, then concatenate accumulated global audio token sequence and decode them to waveform with the DAC codec.

This two-system sliding procedure preserves measure-level alignment across long scores by enforcing a monotonic handoff exactly at the separator. It minimizes boundary artifacts and cumulative tempo drift by reusing the previously generated second-system audio verbatim. The scalability of the length is achieved entirely in inference time through the sliding two-system window; no additional training or architectural change is required.

## III. Datasets

All details are elaborated in [1].

### A. YouTube Score-Video Dataset (YTSV)

We use the large in-the-wild YouTube score–following video corpus introduced in our previous work [1]. We obtain 433,920 image–audio segments drawn from 12,217 videos (1,341 h total). Two videos of *Beethoven: 32 Variations in C minor, WoO 80 – Theme and Variations 1–5* (YouTube IDs `KZiSpxyUsSg` and `nA0cCarOf54`) are present in the training split; none of the other submission pieces appear in the training data.

### B. Additional Modal Datasets

To supply complementary modality supervision, we add three targeted resources:

- **GrandStaff** [9] (augmented synthetic pianoform systems) and **OLiMPiC** [4] (synthetic + scanned pianoform pages) are used for OMR task, enlarging visual engraving diversity and stabilizing symbol decoding.
- **MAESTRO** [10] (199 h paired performance audio + aligned MIDI) is used for the MIDI-to-Audio synthesis task, anchoring high-fidelity timing and pedal articulation in the audio codec token space.

## IV. Post-processing

None applied. We submit the raw waveform obtained by decoding the model-inferenced audio token sequence with the DAC codec. No tempo/dynamics control, symbolic quantization, editing, EQ, reverb, or mastering was performed; expressivity is entirely model-internal.

## References

[1] J. Jung, D. Kim, S. Lee, S. Cho, H. Soh, I. Bukey, C. Donahue, and D. Jeong, "Unified cross-modal translation of score images, symbolic music, and performance audio," 2025.

[2] D. Lee, C. Kim, S. Kim, M. Cho, and W. Han, "Autoregressive image generation using residual quantization," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pp. 11513–11522, IEEE, 2022.

[3] R. Kumar, P. Seetharaman, A. Luebs, I. Kumar, and K. Kumar, "High-fidelity audio compression with improved RVQGAN," in *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023* (A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, eds.), 2023.

[4] J. Mayer, M. Straka, J. Hajič, and P. Pecina, "Practical end-to-end optical music recognition for pianoform music," in *Document Analysis and Recognition - ICDAR 2024* (E. H. Barney Smith, M. Liwicki, and L. Peng, eds.), (Cham), pp. 55–73, Springer Nature Switzerland, 2024.

[5] J. Gardner, I. Simon, E. Manilow, C. Hawthorne, and J. H. Engel, "MT3: multi-task multitrack music transcription," in *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*, OpenReview.net, 2022.

[6] S. Chang, E. Benetos, H. Kirchhoff, and S. Dixon, "Yourmt3+: Multi-instrument music transcription with enhanced transformer architectures and cross-dataset stem augmentation," in *2024 IEEE 34th International Workshop on Machine Learning for Signal Processing (MLSP)*, pp. 1–6, IEEE, 2024.

[7] Musescore, "MuseScore.com — The world's largest free sheet music catalog and community — musescore.com." https://musescore.com/. [Accessed 09-05-2025].

[8] R. Varghese and S. M., "Yolov8: A novel object detection algorithm with enhanced performance and robustness," in *2024 International Conference on Advances in Data Engineering and Intelligent Computing Systems (ADICS)*, pp. 1–6, 2024.

[9] A. Ríos-Vila, D. Rizo, J. M. Iñesta, and J. Calvo-Zaragoza, "End-to-end optical music recognition for pianoform sheet music," *International Journal on Document Analysis and Recognition (IJDAR)*, vol. 26, no. 3, pp. 347–362, 2023.

[10] 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," 2019.