Unified Cross-modal Translation of Score Images, Symbolic Music, and Performance Audio

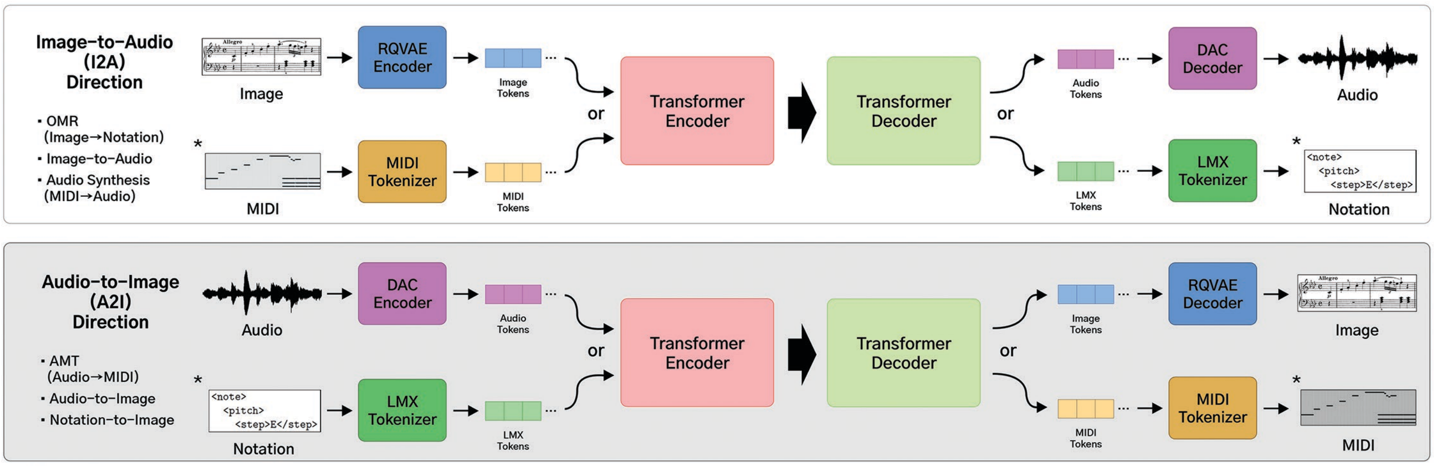
Jongmin Jung[1], Dongmin Kim[1], Sihun Lee[1], Seola Cho[2], Hyungjoon Soh[3], Irmak Bukey[4], Chris Donahue[4], Dasaem Jeong[2]

[1]Department of Artificial Intelligence, Sogang University, Seoul, Republic of Korea, [2]Department of Art & Technology, Sogang University, Seoul, Republic of Korea
[3]Department of Physics Education, Seoul National University, Seoul, Republic of Korea, [4]Computer Science Department, Carnegie Mellon University, Pittsburgh, United States

MALer MUSIC & ARTS LEARNING LAB

- The first successful end-to-end system that converts sheet music images directly into performance audio(I2A).
- State-of-the-art results in core tasks such as Optical Music Recognition (OMR).
- Introduction of the YouTube Score Video (YTSV) dataset, with over 1,300 hours of paired score images and performance audio.

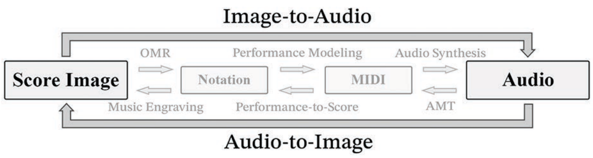
Demo Website

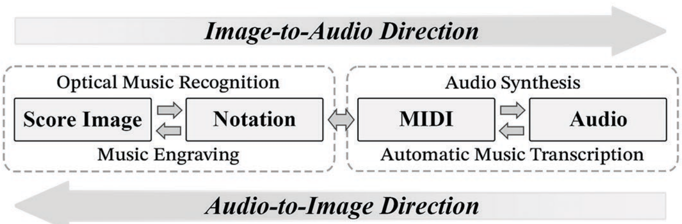


*Already discrete

Overview

- Music can be represented in different formats—score images, music notation (e.g., MusicXML), MIDI, and audio—yet most existing methods focus on only one or two tasks (e.g., Optical Music Recognition, Automatic Music Transcription) separately.
- This research proposes a unified framework approach to **simultaneously learn multiple cross-modal music translation tasks** along the model directions.
- By training **with far-modal translations**, the model implicitly learns to bridge intermediate steps, which **enhances performance on related near-modal tasks**.





Four modalities of music representation in the modal spectrum, along with six cross-modal translation tasks

Model Architecture

- Single seq2seq transformer encoder-decoder with a **unified vocabulary** across image, audio, and symbolic tokens.
- Transformer sub-decoder for decoding the multi-codebook RVQ tokens.
- Two main model directions:

  Image→Audio (I2A): OMR (image→LMX), Performance Audio Synthesis(MIDI-to-audio), and the image-to-audio task.

  Audio→Image (A2I): AMT(audio-to-MIDI), Engraving(LMX-to-image), and the audio-to-image task.

YouTube Score Video (YTSV) Dataset

- 12,217 videos, **433,920 image-audio pairs**, totaling about **1,341 hours** of music.
- Alignment of sheet images (slide by slide) with recorded performance audio.
- Emphasis on classical piano, smaller ensembles (string quartets, etc.), covering a diverse repertoire.

An example of score-following video on YouTube
Slides of sheet music are aligned to the corresponding points in audio. Each systems are then cropped out from the slides.

Category	Videos	Segments	Duration (hrs)
Solo Piano	9,052	232,029	762.34
Accompanied Solo	912	47,373	141.83
String Quartet	594	48,470	138.48
Others (Chamber)	1,659	106,048	298.65
Total	12,217	433,920	1,341

Data distribution of the YouTube Score Video dataset

Evaluation Method & Results

◦ Optical Music Recognition(OMR)

- Symbol Error Rate (SER) metric on LMX token sequences.[1]

Method	OLiMPiC		BPSD
	Synth	Scanned	Scanned
OMR-only	15.90	24.58	45.39
+ Image-to-Audio	10.57	15.45	23.85
+ MIDI-to-Audio	**9.72**	**13.67**	**23.36**
Zeus	10.10	14.45	31.24

OMR Results in SER, compared to Zeus[1]. Lower is better.

◦ Automatic Music Transcription(AMT)

- Note-F_1 score implemented in the `mir-eval` library.[2]

Method	MusicNet$_{EM}$		MAESTRO
	Str	WW	
AMT-only	87.21	72.04	89.40
+ Audio-to-Image	**87.28**	72.61	89.38
+ LMX-to-Image	87.25	**75.52**	**89.45**
Maman *et al.* [39]	80.0	87.5	89.7
Chang *et al.* [18]	91.32	83.46	96.98

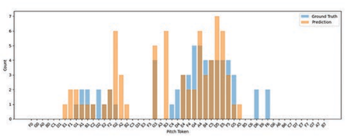
AMT results in note onset F_1 score for test set, compared to Maman et al. (2022)[3]. Higher is better.

◦ Audio-to-Image(A2I)

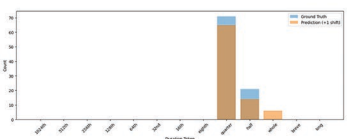
- Evaluate the results by comparing LMX token histograms from generated and ground truth scores using Earth Mover's Distance (EMD).
- Compute EMD separately for pitch and duration tokens, with ±1 shift for durations to account for meter interpretation differences.

Method	EMD ↓	
	Pitch	Duration
Audio-to-Image Only	4.6436	0.4873
+ AMT	2.8880	0.4377
+ LMX-to-Image	**2.6350**	**0.4317**
GT Random Pairing Baseline	3.4921	0.9936
RQVAE Reconstruction	0.8990	0.1301
GT Image	0.4865	0.1113

Audio-to-image generation accuracy in EMD on BPSD



Histograms of prediction/GT pitch tokens



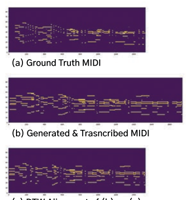
Histograms of prediction/GT duration tokens

◦ MIDI-to-Audio(M2A; Performance Audio Synthesis)

- **Transcribe the output audio** with Onsets and Frames model[4] and calculate **Note-F_1** score on 3 different onset thresholds(50, 100, 200ms).
- Frechet Audio Distance(FAD)[5] as a general audio quality metric.

Method	F_1 ↑			FAD ↓
	50ms	100ms	200ms	
MIDI-to-Audio Only	26.61	64.86	**88.20**	0.201
+ OMR + I2A	**39.37**	**66.63**	84.66	**0.143**
MSD [51]	83.82	90.63	92.28	0.229

MIDI-to-audio synthesis accuracy in F_1 and FAD on BPSD



(a) Ground Truth MIDI
(b) Generated & Trasncribed MIDI
(c) DTW Alignment of (b) on (a)

◦ Image-to-Audio(I2A)

- Calculate Note-F_1 score and FAD like in M2A, but apply **Dynamic Time-Warping(DTW)** between the transcribed MIDI and the ground truth MIDI, before calculating Note-F_1 score.
- In the "multi stage" strategies (Image → MusicXML → MIDI → Audio), pre-trained models to perform each translation step as a baseline: Zeus[1] for OMR, VirtuosoNet[6] for performance modeling, and Music Spectrogram Diffusion[7] for audio synthesis.

Method (Metric / Dataset / Onset Tolerance / Criteria)	F_1 Score ↑ BPSD 50ms	BPSD 100ms	BPSD 200ms	YTSV-T11 50ms	YTSV-T11 100ms	YTSV-T11 200ms	FAD ↓ BPSD —	FAD ↓ YTSV-T11 —	MOS ↑ BPSD-T12 NF	PN	AQ
Direct I2A: YTSV-P (I2A Only)	23.49	34.51	44.15	27.05	43.32	53.02	0.422	0.317	1.26	1.52	2.05
Direct I2A: OMR + I2A	48.67	64.30	74.01	51.60	67.92	75.98	0.098	0.056	—	—	—
Direct I2A: OMR + I2A + M2A	48.36	64.63	74.92	**52.66**	**68.45**	**76.24**	**0.081**	**0.055**	**3.92**	**3.51**	**3.50**
Multi-stage: OMR + I2A+ M2A	**50.91**	**70.40**	**79.96**	—	—	—	0.137	–	3.21	2.75	3.07
Multi-stage: Zeus → VNet → MSD	45.52	59.35	69.36	—	—	—	0.330	—	2.37	2.21	2.66
DAC Reconstruction (Upper-bound)	68.83	82.39	87.47	82.28	86.43	88.76	0.050	0.035	—	—	—
Ground Truth (MOS Upper-bound)	—	—	—	—	—	—	—	—	4.80	4.68	4.24

Accuracy of image-to-audio generation in terms of note onset F_1 score and FAD

References
[1] Mayer, J., Straka, M., Haji ˇc, J., and Pecina, P. Practical end-to-end optical music recognition for pianoform music. InBarney Smith, E. H., Liwicki, M., and Peng, L. (eds.), Document Analysis and Recognition - ICDAR 2024, pp. 55–73, Cham, 2024. Springer Nature Switzerland. ISBN978-3-031-70552-6.
[2] Raffel, C., McFee, B., Humphrey, E. J., Salamon, J., Nieto, O., Liang, D., Ellis, D. P., and Raffel, C. C. Mir eval: A transparent implementation of common mir metrics. In ISMIR, volume 10, pp. 2014, 2014.
[3] Maman, B. and Bermano, A. H. Unaligned supervision for automatic music transcription in the wild. In Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., and Sabato, S. (eds.), Proceedings of the 39th International Conference on Machine Learning, volume 162 of Proceedings of Machine Learning Research, pp. 14918–14934. PMLR, 17–23 Jul 2022.
[4] Hawthorne, C., Elsen, E., Song, J., Roberts, A., Simon, I., Raffel, C., Engel, J., Oore, S., and Eck, D. Onsets and frames: Dual-objective piano transcription. arXiv preprint arXiv:1710.11153, 2017.
[5] Azalea Gui, Hannes Gamper, S. B. D. E. Adapting frechet audio distance for generative music evaluation. In Proc. IEEE ICASSP 2024, 2024.
[6] Jeong, D., Kwon, T., Kim, Y., and Nam, J. Graph neural network for music score data and modeling expressive piano performance. In International conference on machine learning, pp. 3060–3070. PMLR, 2019.
[7] Hawthorne, C., Simon, I., Roberts, A., Zeghidour, N., Gardner, J., Manilow, E., & Engel, J. (2022). Multi-instrument Music Synthesis with Spectrogram Diffusion. International Society for Music Information Retrieval Conference.