U-MusT: A Unified Framework for Cross-modal Translation of Score Images, Symbolic Music, and Performance Audio

Jongmin Jung*, Dongmin Kim*, Sihun Lee, Seola Cho, Hyungjoon So, Irmak Bukey, Chris Donahue, and Dasaem Jeong†
*These two authors contributed equally.

***Abstract*—Music exists in various modalities, such as score images, symbolic scores, MIDI, and audio. Translations between such modalities are established as core tasks of music information retrieval, such as automatic music transcription (audio-to-MIDI) and optical music recognition (score image to symbolic score). However, most past work on multimodal translation utilizes specialized models trained for each translation task. In this paper, we propose a unified framework based on a common tokenization strategy. We use dedicated separate models for the Image-to-Audio and Audio-to-Image directions, sharing an identical encoder-decoder architecture to handle each task within a coherent framework. Two key factors make this unified approach viable: a new large-scale dataset, and the tokenization of each modality. Firstly, we propose a new dataset that consists of more than 1,300 hours of paired audio-score image data collected from YouTube videos, which is an order of magnitude larger than any existing music modal translation datasets. Secondly, our unified tokenization framework discretizes score images, audio, MIDI, and MusicXML into a sequence of tokens, enabling standard encoder–decoder Transformers to tackle multiple cross-modal translation as one coherent sequence-to-sequence task. Experimental results confirm that our unified framework improves upon single-task baselines in several key areas, notably reducing the symbol error rate for optical music recognition from 24.58% to a state-of-the-art 13.67%, while also seeing substantial improvements across the other translation tasks. Notably, our approach achieves the first musically-coherent score-image-conditioned audio generation, marking a significant breakthrough in cross-modal music generation.**

***Index Terms*—Cross-modal music translation, Multi-task Learning, Optical music recognition, Automatic music transcription, Image-to-audio, MIDI-to-audio, Music information retrieval, YouTube Score Video dataset.**

I. Introduction

A piece of music can be represented in various forms, each with distinct characteristics. For example, a Beethoven piano sonata would typically be distributed as printed scores, where the composer notates their musical ideas with symbols; these scores may exist as scanned or synthesized images, or in machine-readable notation formats such as MusicXML. Specific performances of the piece can be recorded as audio files but can also be represented in time-aligned symbolic formats such as MIDI, where each note's onset and offset timing are explicitly provided.

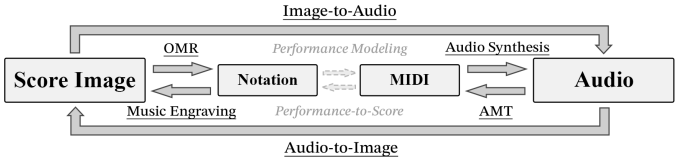

Fig. 1. The four modalities of music representation in the *modal spectrum*, along with six cross-modal translation tasks tackled in this paper (underlined). Far-modal translation tasks, such as *Image-to-Audio* and *Audio-to-Image*, inherently necessitate the implicit resolution of intermediate near-modal tasks (e.g., OMR, AMT, performance modeling and performance-to-score).

As each modality serves a different purpose, converting music from one representation to another has garnered much research interest in the field of music information retrieval (MIR) and established numerous mainstay tasks: as illustrated in Fig. 1, automatic music transcription (AMT) [1], [2], MIDI-to-audio synthesis [3], [4], optical music recognition (OMR) [5], [6], complete music transcription (audio-to-notation) [7], performance modeling [8], and performance-to-score conversion [9], [10].

However, previous research has largely treated these tasks as separate problems, relying on specialized datasets and methodologies designed for each task. Some works have explored multimodal pipelines—first converting music notation into performance MIDI and subsequently synthesizing audio [11], [12]—but typically through multi-stage frameworks using separate models.

We can instead consider a task that inherently encompasses multiple modal translations at once. The most extreme case of this would be score image to performance audio conversion—— the process of directly generating music audio from a score image, bypassing symbolic notation altogether. To the best of our knowledge, this task has never been attempted before, likely because it demands mastery of both OMR and controllable music generation, two open areas of research. Yet, it mirrors how human musicians interpret music: when sight-reading, performers directly *translate* score images into expressive performances, without needing an intermediate symbolic representation.

In this paper, we propose a methodology for multimodal music translation, based on a new large-scale dataset and a

Jongmin Jung, Dongmin Kim, Sihun Lee are with Department of Artificial Intelligence, Sogang University, Seoul, South Korea. Seola Cho is with Sogang Future Lab, Sogang University, Seoul, South Korea. Hyungjoon So is with Department of Physics Education, Seoul National University, Seoul, South Korea. Irmak Bukey and Chris Donahue are with Computer Science Department, Carnegie Mellon University, Pittsburgh, PA, United States. Dasaem Jeong is with Department of Art & Technology, Sogang University, Seoul, South Korea.
†Corresponding author: dasaemj@sogang.ac.kr

This work was supported by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (NRF-2024S1A5C3A03046168).

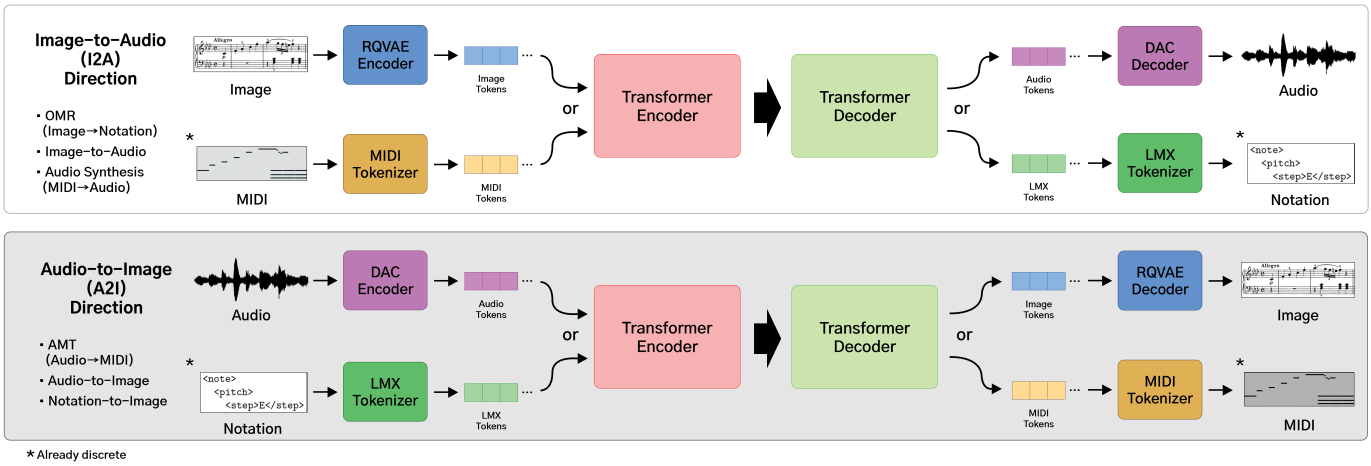

Fig. 2. Overview of our proposed methodology for multimodal music translation. We employ a single Transformer encoder-decoder model for each direction—one for *Image-to-Audio direction* (I2A) tasks and another for *Audio-to-Image direction* (A2I) tasks. Each model jointly handles multiple translation tasks. All modalities are discretised into token sequences, enabling end-to-end, multi-task training entirely at the token level. Note that we train separate models for I2A and A2I directions; the two directions do not share weights.

shared sequence-modeling paradigm. Following the previous research in unifying image-language-speech modal translation [13], we formulate music modal translation as a seq2seq task using vector-quantized discrete tokens of image and audio. This coherent formulation enables a synergistic interaction between different modal translation tasks, where information learned from one task can benefit others, even when the input and output modalities do not directly overlap. Specifically, our result shows that OMR performance can be improved when trained together with MIDI-to-audio synthesis.

While the image-audio pairs in our new dataset do not include any symbolic information such as MusicXML or MIDI, their musical semantics are closely connected. As illustrated in Fig. 1, to generate audio from a given score image, a model must implicitly handle all modal translations: recognizing notes and instructions from score images, predicting the timing and dynamics of each note, and synthesizing the notes into audio. Conversely, for audio-to-image generation, the model must identify the notes from audio, organize them into formalized notation, and engrave the notation in sheet music form. Based on our experiments, we show that learning *outer* modality translations (image $\leftrightarrow$ audio) can also enhance performance for *inner* modality tasks such as OMR and AMT.

Furthermore, we find that our image-to-audio and audio-to-image translation models function properly only when trained alongside other modal tasks that align with their translation direction (as illustrated in Fig. 2). This highlights that the unification of modal translation tasks is not just an optional enhancement but a necessary component in making score-to-audio or audio-to-score generation feasible.

Our main contributions are as follows:

- We propose the first unified multimodal music translation methodology that covers the four most representative modalities of Western music—score images, symbolic scores, performance MIDI, and audio—within a shared architectural framework.
- We are the first to train a model that can generate musically-coherent audio given a score image, closing the gap with the capabilities of human musicians.
- We introduce the YouTube Score Video (YTSV) dataset that consists of more than 1,300 hours of score images paired with corresponding performance audio recordings.
- We demonstrate that the combined multimodal training can enhance the performance of subtasks.

II. Problem Formulation and Related Work

Before detailing our methodology, it is essential to clarify the distinct modalities of music representation that our framework aims to bridge. In this work, we consider the following four modalities of music representation:

- **Score Image**: Sheet music in the form of raw images. Typically scanned from real-world prints or rendered with digital notation software. We cover both synthetic and scanned images while discarding hand-written scores.
- **Symbolic Notation**: Digitized semantic information of sheet music, such as MusicXML (MXL). Along with the pitch and duration of notes, it also includes detailed elements such as slurs, voicing, articulation, and tempo and dynamic markings that are essential to render human-readable music scores. We employ a modified version of the industry-standard format MusicXML, which we describe further in Section III-B.
- **MIDI**: Event-based symbolic representation of the performed notes with concrete note onset and offset timings. While notational details such as slurs or articulation are omitted, MIDI is better suited to express human performance with expressive tempo and dynamic changes.
- **Audio**: Recorded from real-world performance or synthesized with virtual instruments.

As shown in Fig. 1, we observe that these four representations exist on a *modal spectrum*; each one naturally leads to the next in the process of music-making and in MIR tasks. As the modalities have a strong causal relationship with one another and each highlight different aspects of music, the information

learned for one subtask can have a synergistic effect with others. We consider translation tasks starting from the score image-side as being in the *Image-to-Audio direction* (**I2A**), and translation tasks starting from the audio-side as being in the *Audio-to-Image direction* (**A2I**).

Among many possible modal translations, we mainly focus on automatic music transcription and optical music recognition, the key representative tasks in each direction that are actively researched. In addition, we also introduce direct image-to-audio and audio-to-image translation as novel challenging tasks that expand the boundaries of cross-modal music translation beyond adjacent conversions, addressing the end-to-end capabilities of translation models.

A. Automatic Music Transcription

Automatic music transcription (AMT) is the task of inferring musical notes or notations from audio, and is considered one of the fundamental problems in MIR. The output of AMT can range anywhere from simple piano-roll representations to structured sheet music data [14]. Machine learning-based AMT has seen much success in recent years, particularly with piano music [15], [16], thanks to the vast amount of available training data with high precision labels collected with computer-controlled pianos [3]. While piano-roll-like frame-wise prediction has been widely used for AMT models, Hawthorne *et al.* [17] introduced the first AMT model utilizing token-based prediction using a transformer encoder-decoder structure. This approach has been further adapted to multi-instrument AMT [18], [19]. Beyond direct transcription, the related task of audio-to-symbolic arrangement has been addressed through cross-modal representation learning, where a variational autoencoder learns to disentangle musical features such as chords and texture from an audio input to generate a symbolic arrangement [20].

However, extending these successes to classical ensembles beyond the piano remains challenging due to the scarcity of synchronized audio-MIDI training labels. To address this limitation, Maman *et al.* [21] proposed a method to leverage score-derived MIDI (which possess relative or metrical timing) and their corresponding audio recordings. They utilized Dynamic Time Warping (DTW) to align the score-based MIDI with the MIDI transcribed directly from the audio, effectively synchronizing the symbolic data along the time axis to augment the ground truth for training.

B. Optical Music Recognition

The aim of optical music recognition (OMR) is to transcribe music notation from score images. While previous methods typically involve multi-stage processes, recent state-of-the-art approaches adopt an end-to-end strategy that takes images as input and predicts token sequences for the corresponding symbolic notation [22], [23]. Due to the difficulty in handling complex score layouts, it was only recently that OMR research on polyphonic piano form score images has seen successful results [24].

A significant bottleneck in OMR research is the limited availability of labeled training data in various styles; while many resources exist for machine-readable symbolic music notation, corresponding real-world scans are relatively scarce. Therefore, much of OMR research relies on synthetic datasets for model training, while reserving scanned images for evaluation [23], [25].

C. Generative and Conversion Tasks: Synthesis, Modeling, and Rendering

Beyond transcription, generating and converting musical content across symbolic and audio representations constitutes another major domain. MIDI-to-audio synthesis has been explored using various neural approaches [3], [4], which generate high-fidelity audio from MIDI conditioning. Concurrently, performance modeling aims to imbue MIDI sequences with human-like expressivity; models like VirtuosoNet [8] predict expressive performance characteristics such as tempo and dynamics from score information. Conversely, the task of performance-to-score conversion addresses the inverse challenge, recovering structured notation from expressive performance MIDI [9], [10].

In the visual domain, the task of generating score images—often referred to as music engraving—has traditionally been handled by rule-based software like MuseScore [26], leaving little practical demand for deep learning-based solutions. While recent work such as MusicScore [27] has attempted to generate score images using latent diffusion models, evaluations have primarily focused on image quality metrics like FID [28], often neglecting critical assessments of musical semantic accuracy and readability as functional sheet music. However, in the context of Optical Music Recognition (OMR), synthetic score generation serves a critical role in mitigating the scarcity of annotated handwritten data. Shatri *et al.* [29] presented a comprehensive evaluation of GAN architectures such as DCGAN, ProGAN, and CycleWGAN for this purpose. They demonstrated that CycleWGAN, enhanced with Wasserstein loss for improved training stability, achieves superior performance in style transfer, generating diverse and high-quality synthetic handwritten sheets from printed inputs to augment OMR training datasets.

D. Multimodal and Multi-task Approaches

Mitigating the interference between diverse tasks while maintaining a unified backbone is a central challenge in multimodal learning. To address this, approaches like Mixpert [30] and TaskExpert [31] have adopted Mixture-of-Experts (MoE) architectures to dynamically adapt to diverse domains. Additionally, methods such as Self-MM [32] show that self-supervised multi-task learning can successfully isolate modality-specific characteristics without costly labeling.

Narrowing down to the audio and music domain, recent works have explored unified models that tackle multiple tasks and modalities simultaneously. UniAudio [33] introduced a single model trained on 11 diverse audio generation tasks, spanning inputs from audio, text, MIDI, and phonetic sequences, by leveraging a discrete audio token representation. Fugatto [34] likewise demonstrated that a unified model conditioned on both audio and text can achieve remarkably

fluent results across various audio generation tasks. Another notable attempt was proposed by Kim *et al.* [13], which trains a single sequence-to-sequence Transformer on discrete token sequences to translate between images, text, and speech. Crucially, recent research has begun to challenge the necessity of fine-grained supervision in music tasks. Zeng *et al.* [35] demonstrated that joint expressive rendering and transcription can be achieved using only sequence-aligned data. However, these prior works predominantly operate on symbolic representations or well-aligned datasets, leaving the complex interplay between *raw visual* scores and audio performance largely unexplored.

Bridging this gap, our work extends the unified multimodal paradigm to the domain of music translation across score images and performance audio. We introduce new translation tasks underpinned by a novel large-scale dataset and a learning approach designed to overcome imperfect alignments. In a related vein, the principle of leveraging multimodal data has also been explored through cross-modal knowledge distillation, where a richer audio-visual model enhances an audio-only speech separation system [36], demonstrating the broad applicability of cross-modal learning paradigms. Specifically, our proposed model handles a broader set of output modalities beyond audio tokens: in addition to the traditional tasks of automatic music transcription and optical music recognition, we enable cross-modal generation of performance audio from sheet music images, and vice versa (generating sheet images from audio). To mitigate the data scarcity in tasks like AMT and OMR, we contribute a new large-scale dataset (over 1,300 hours of sequence-aligned score-image/audio pairs). Leveraging the potential of sequence-level supervision, even though these image–audio pairs lack explicit note-level alignments, we demonstrate that incorporating them into a unified multi-task framework allows the model to implicitly learn shared musical structure, leading to improved performance on transcription tasks.

III. Unified Tokenization Framework

To enable our unified sequence-to-sequence modeling across diverse music modalities, we convert every input and output into a sequence of discrete tokens. We denote a *modality* by the calligraphic symbol $\mathcal{X} \in \{\mathcal{I}, \mathcal{A}, \mathcal{N}, \mathcal{M}\}$ for **I**mage, **A**udio, musical **N**otation (LMX), and **M**IDI performance, respectively. The corresponding *raw* data X are written with roman capitals I, A, N, M. Each modality owns a *tokenizer encoder* $\mathcal{F}_{\mathcal{X}}$ and an inverse *tokenizer decoder* $\mathcal{G}_{\mathcal{X}}$ such that

$$\mathcal{F}_{\mathcal{X}}(X) = z^{(\mathcal{X})}_{1:L_X}, \qquad \mathcal{G}_{\mathcal{X}}\big(z^{(\mathcal{X})}_{1:L_X}\big) \approx X. \tag{1}$$

Hence, $z^{(\mathcal{X})}_{1:L_X}$ is the discrete-token representation of X and L_X is its length (we subsequently write L_I, L_A, L_N, L_M for the four modalities). Each modality has its own vocabulary $\mathcal{V}_{\mathcal{X}}$, yet all tokens ultimately live in a unified vocabulary $\mathcal{V}$ (Section III-D), allowing a single Transformer to translate between any pair of modalities. Continuous data modalities such as score images and audio are discretized using learned neural compression models, which quantize the raw image pixels or audio waveform into compact token sequences. Symbolic modalities such as MusicXML scores and MIDI performances are tokenized by structurally flattening their hierarchical representations into linear sequences of musical tokens. This unified tokenization scheme provides a common sequential format for all modalities and is essential for training a single Transformer that can handle the full range of cross-modal music translation tasks.

A. Image ($\mathcal{X} = \mathcal{I}$) and Audio ($\mathcal{X} = \mathcal{A}$) Tokens

We train two discrete representation models: Residual Quantized Variational Autoencoder (RQVAE) [37] and Descript Audio Codec (DAC) [38] for images and audio, respectively. Both tokenizers employ $d = 4$ *unshared* codebooks, each of cardinality $\kappa = 1024$. Consequently, each time-step yields a *bundle* of d code indices $z_{t,1}, \ldots, z_{t,d} \in \{0, \ldots, \kappa{-}1\}$, so the discrete representation is a 2-D array of shape $L_X \times d$. We still refer to its "length" as L_X for notational simplicity.

1) Image Tokens: We train RQVAE with compression rate (C) of 16 and a model dimension of 256. Since the original RQVAE architecture was designed for three-channel RGB images, we adapt it to process single-channel grayscale sheet music. While the original model achieved 32× compression, we implement a modified channel multiplication sequence of `[1, 1, 2, 2, 4]` for 16× compression. This change is crucial as it ensures each token captures features at a sub-staff-line scale, which is necessary for a precise representation of musical notation, as illustrated in Fig. 3.

We removed the model's attention blocks to specialize it for local feature extraction. While attention is typically used for global coherence, our task requires tokenizing fine-grained details like note heads, making the attention mechanism unnecessary and potentially detrimental.

Sheet music presents a unique challenge with its varying image dimensions. To address the original model's limitation of being trained on fixed-resolution images, which led to reconstruction artifacts when processing inputs of different sizes (Fig. 4), we implement a dynamic, resolution-adaptive training strategy. The training process uses different random crop and batch sizes based on the input image's height:

- 70˜130 pixels: 64-pixel random crops, batch size 256
- 130˜260 pixels: 128-pixel random crops, batch size 128
- 260˜360 pixels: 256-pixel random crops, batch size 32
- 360˜390 pixels: 352-pixel random crops, batch size 16

We implement a vertical flattening method for 2D image token sequences to match how sheet music is naturally read; The model first processes a vertical column of image (top-to-bottom), then moves horizontally to the next column on the right, mirroring standard sheet music reading patterns (left-to-right). Prior to tokenization, we threshold the images by identifying the median pixel value of each image and setting all pixels above $median - 20$ to white (255), reducing noise and enhancing contrast.

The tokenization process for each musical system can be broken down as follows:
RQVAE compresses each image patch by a factor of $C = 16$.

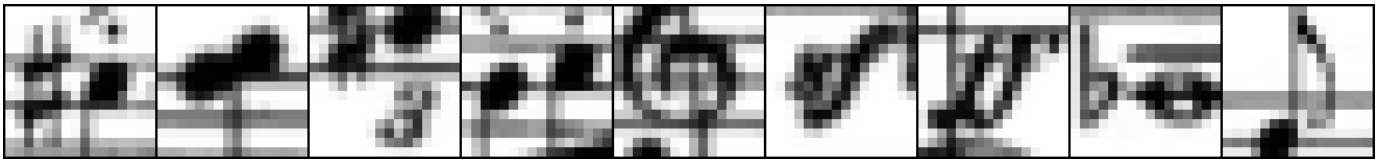

Fig. 3. Example patches showing the 16 × 16 pixel resolution of individual tokens. Each patch captures sub-staff elements, spanning an area of 1-2 note heads.

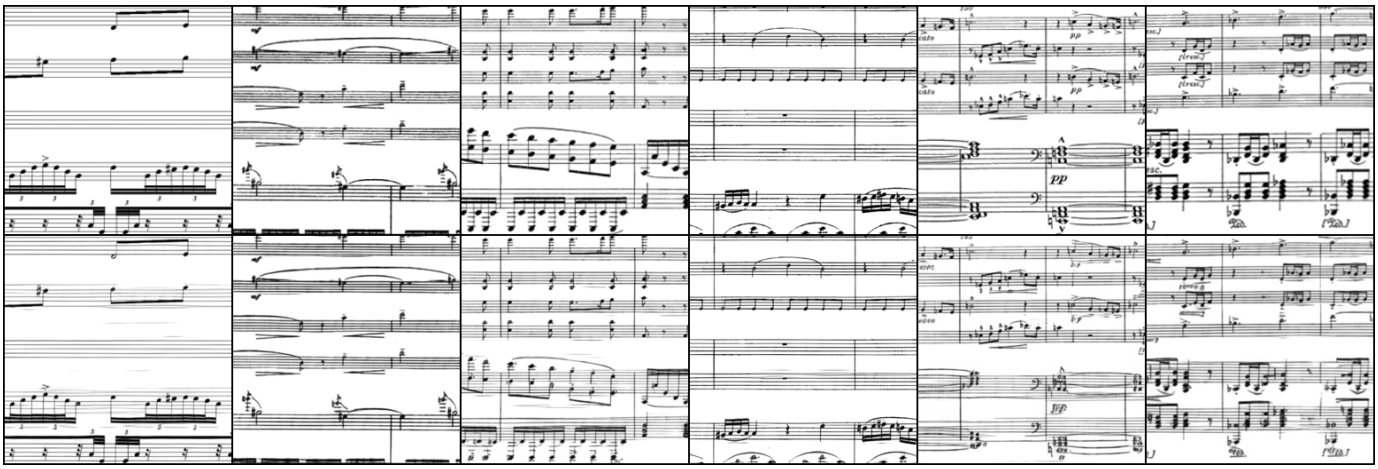

Fig. 4. Comparison between input sheet music (top) and model reconstruction (bottom), demonstrating reconstruction artifacts in staff lines when a model trained only on 64 × 64 pixel image crops processes 256 × 256 pixel inputs(unseen image size).

Given a score image that contains K musical *systems* (i.e. lines of sheet music) $\text{system}_i \in \mathbb{R}^{W_i \times H_i}$, RQVAE produces

$$\text{system}_{i,\text{tokens}} \in \left(\mathcal{V}_\mathcal{I}\right)^{\left(\frac{W_i H_i}{C^2} \times d\right)}. \tag{2}$$

We flatten each $\frac{W_i H_i}{C^2} \times d$ token grid in *vertical* reading order (top-to-bottom inside a column, then left-to-right between columns) and concatenate all systems, inserting the separator token `[SEP]`:

$$\text{image}_{\text{tokens}} = \text{Concat}\big(\text{system}_{1,\text{tokens}},\ \texttt{[SEP]},\ \ldots,\ \texttt{[SEP]},\ \text{system}_{K,\text{tokens}}\big). \tag{3}$$

$$\left|\text{system}_{i,\text{tokens}}\right| = \frac{W_i \times H_i}{C^2} \times d, \tag{4}$$

$$L_I = \sum_{i=1}^{K} \left|\text{system}_{i,\text{tokens}}\right| + (K-1). \tag{5}$$

The `[SEP]` token enables us to process sequences of variable-height system images without normalizing them to a fixed size via padding. This design teaches the model to handle components with non-uniform dimensions, leading to greater memory efficiency. During generation, this allows the RQVAE to reconstruct each `[SEP]`-separated output into an image of its respective, variable height.

2) Audio Tokens: We train DAC with a token hop size h of 512 samples on our 44.1kHz (f_s) source material, resulting in approximately 86.13 RVQ token sets per second ($\frac{f_s}{h}$). The model employs an embedding size of 1,024. We retrain DAC because the publicly available model uses nine codebooks, which is unnecessarily large for our domain. By limiting to four codebooks and training specifically on classical music, we aim to achieve richer representations within a narrower domain while maintaining high reconstruction quality.

All audio is resampled to $f_s = 44.1\,\text{kHz}$ (mono) and tokenized with DAC using hop size $h = 512$ samples. The number of hop frames in T seconds equals $\left\lceil \frac{T f_s}{h} \right\rceil$, giving

$$L_A = \left\lceil \frac{T f_s}{h} \right\rceil, \qquad \text{audio}_{\text{tokens}} \in \left(\mathcal{V}_\mathcal{A}\right)^{L_A \times d}. \tag{6}$$

3) Augmenting Data with Spatial and Temporal Shifts: During tokenization, we observe that even slight shifts in pixel positions or audio samples can result in vastly different token assignments, a limitation inherent to discrete token representations. To address this and enhance model robustness, we implement comprehensive augmentation strategies: for images, we generate 32 variants through combinations of 8 horizontal and 4 vertical pixel shifts, and for audio, we create 9 variations through temporal shifts ranging from -20 to +20 samples at five-sample intervals, with each five-sample shift corresponding to approximately 0.113 milliseconds. More details are elaborated in Appendix B.

B. Linearized MusicXML ($\mathcal{X} = \mathcal{N}$)

MusicXML is a machine-readable musical notation format for storing semantic information of sheet music. While the format's versatility has led to a wide adoption in editing software and music processing libraries, its hierarchical structure and verbosity pose challenges for processing with language models. To mitigate this, Mayer *et al.* [23] introduced *Linearized MusicXML (LMX)*, a modification of the format that is

much more concise and a better fit for token-based generation. We use the original implementation of LMX to represent the notation data in all of our experiments.

Let L_N be the sequence length; the token string is

$$\text{lmx}_{\text{tokens}} \in \big(\mathcal{V}_{\mathcal{N}}\big)^{L_N \times 1}, \tag{7}$$

where by convention $d = 1$. For training and inference we *pad* codebook positions 2–4 with a special `[PAD]` token: $z^{(\mathcal{N})}_{t,2:4} =$ `[PAD]`.

C. MIDI-Like Tokens ($\mathcal{X} = \mathcal{M}$)

MIDI represents musical performance as a sequence of digital events, capturing note information and timing. While this format effectively encodes performance data, its event-based structure requires adaptation for language model processing. Drawing from the MT3 framework [18], we adopt a MIDI-like tokenization scheme that transforms MIDI data into discrete tokens that include instrument identifiers, MIDI pitches, note on/off events, and quantized time markers at 10ms intervals. For this, we use the specific implementation introduced in YourMT3+ [19].

If L_M denotes its length,

$$\text{midi}_{\text{tokens}} \in \big(\mathcal{V}_{\mathcal{M}}\big)^{L_M \times 1}, \tag{8}$$

and again $z^{(\mathcal{M})}_{t,2:4} =$ `[PAD]`.

D. Unified Vocabulary ($\mathcal{V}$)

The overall vocabulary $\mathcal{V}$ is the union of all modality-specific token sets and these special tokens.

$$\mathcal{V} = \Big(\bigcup_{\mathcal{X}} \mathcal{V}_{\mathcal{X}}\Big) \cup \Big(\bigcup_{\mathcal{X}} \{\texttt{[SOS]}_{\mathcal{X}}, \texttt{[EOS]}_{\mathcal{X}}\}\Big) \cup \{\texttt{[SEP]}, \texttt{[PAD]}\}. \tag{9}$$

IV. Model Architecture

A single sequence-to-sequence encoder-decoder Transformer [39] model is used to perform the translation between token sequences with different modalities. The encoder $\mathcal{E}$ takes as input the source token sequence $z^{(\mathcal{X})}_{1:L_X}$ (of length L_X) and maps it to a sequence of hidden representations; the decoder $\mathcal{D}$ autoregressively generates the target sequence $\hat{z}^{(\mathcal{Y})}_{1:L_Y}$ token by token. In a glimpse, given a source modality $\mathcal{X}$ and target modality $\mathcal{Y}$,

$$z^{(\mathcal{X})}_{1:L_X} = \mathcal{F}_{\mathcal{X}}(X), \qquad H = \mathcal{E}\big(z^{(\mathcal{X})}_{1:L_X}\big), \qquad \hat{z}^{(\mathcal{Y})}_{1:L_Y} = \mathcal{D}(H), \tag{10}$$

and finally $\hat{Y} = \mathcal{G}_{\mathcal{Y}}\big(\hat{z}^{(\mathcal{Y})}_{1:L_Y}\big)$. To inform the model of the desired target modality, we add a target-modality embedding $\text{TgtEmb}_{\mathcal{Y}}$ to every input token representation. Also, each modality has a modality-specific learnable positional embedding $\text{PosEmb}_{\mathcal{X}}$. Thus, each source token is embedded as

$$e_i = \text{TokEmb}(z^{(\mathcal{X})}_i) + \text{PosEmb}_{\mathcal{X}}(i) + \text{TgtEmb}_{\mathcal{Y}}. \tag{11}$$

Images and audio emerge from *bundles* of d=4 RVQ code-tokens per time step, whereas symbolic modalities emit only a single token. If we forced the Transformer decoder to emit the entire *flattened* code-token stream, the effective sequence length would be inflated by $\times d$, yielding slower training, poorer long-range attention, and an interface mismatch with symbolic tasks. Instead, we follow the codebook-wise decoding strategy of Yang *et al.* [33]. We keep the main decoder autoregressive over *temporal and spatial musical time* and delegate the intra-step codebook prediction to a lightweight *sub-decoder* $\mathcal{D}_{sub}$, achieving consistent sequence lengths and uniform decoding logic across all modalities. The main decoder $\mathcal{D}_{main}$ emits a hidden state $h_t \in \mathbb{R}^D$ for each *time step* t. To transform h_t into d code tokens we introduce a one-layer Transformer *sub-decoder* $\mathcal{D}_{sub}$:

$$(z_{t,1}, \ldots, z_{t,d}) = \mathcal{D}_{sub}\big(h_t\big), \quad h_t = \mathcal{D}_{main}\big(H\big). \tag{12}$$

- **Input to $\mathcal{D}_{sub}$.** For sub-step ℓ ($1 \le \ell \le d$) we sum (i) the token embedding of the previous output $\hat{z}_{t,\ell-1}$ (or $\texttt{SOS}_{sub}$ if $\ell = 1$), (ii) a dedicated positional embedding for ℓ, and (iii) the broadcast main-decoder state h_t: $e_{t,\ell} = \text{TokEmb}_{sub}(\hat{z}_{t,\ell-1}) + \text{PosEmb}_{sub}(\ell) + h_t$.
- **Symbolic targets.** If $\mathcal{Y} \in \{\mathcal{N}, \mathcal{M}\}$ we set $d = 1$; codebooks $2-4$ are always `[PAD]`.

We train the model using the standard cross-entropy loss to maximize the likelihood of the target token sequence given the source sequence. With cross-entropy loss $\text{CE}(\cdot,\cdot)$ and parameter set θ,

$$\mathcal{L}(\theta) = -\sum_{t=1}^{L_Y} \begin{cases} \sum_{\ell=1}^{d} \log P_\theta\big(z^{(\text{gt})}_{t,\ell} \mid z^{(\text{gt})}_{<t,*}, H\big), & \mathcal{Y} \in \{\mathcal{I}, \mathcal{A}\} \\ \log P_\theta\big(z^{(\text{gt})}_{t,1} \mid z^{(\text{gt})}_{<t,1}, H\big), & \mathcal{Y} \in \{\mathcal{N}, \mathcal{M}\}. \end{cases} \tag{13}$$

When applying $softmax$ for each modality, entries of other modalities are masked as $-inf$. During training we employ teacher-forcing, and at inference, the model produces outputs autoregressively. This unified model design allows a single Transformer to learn all modal translations jointly under a consistent tokenization scheme.

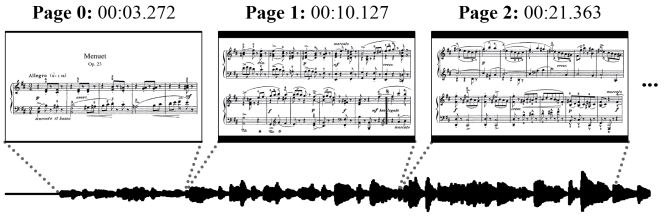

Fig. 5. An example from one of the videos collected for the YouTube Score Video dataset. Slides of sheet music are aligned to the corresponding points in audio.

V. YouTube Score Video Dataset

The primary advantage of the proposed unified, end-to-end framework is its ability to leverage vast quantities of paired data, even across distant modalities such as score images and performance audio (Fig. 1). While such image-audio pairs are not an essential condition for near-modal translation tasks like OMR or AMT, with enough volume they present an opportunity for synergistic learning. As posited in Section I, the task of translating a score image directly to audio requires implicit handling of intermediate near-modal

TABLE I
DATA DISTRIBUTION OF THE YOUTUBE SCORE VIDEO DATASET AFTER FILTERING.

Category	Description	Videos	Segments	Duration (hrs)
Piano Solo	Solo piano compositions	9052	232029	762.34
Accompanied Solo	Solo compositions for a non-piano instrument with piano accompaniment	912	47373	141.83
String Quartet	Compositions for two violins, viola, and cello	594	48470	138.48
Others	Compositions not classified under predefined categories	454	24912	69.13
Unaccompanied Solo	Solo compositions for a single non-piano instrument	207	3542	11.24
Guitar Solo	Solo compositions for classical guitar	192	1976	6.97
Piano Trio	Compositions for piano, violin, and cello	254	22736	68.51
Organ Solo	Solo compositions for organ	161	5923	20.01
Piano Quintet	Compositions for piano and string quartet	109	13382	34.69
Piano Quartet	Compositions for piano, violin, viola, and cello	84	9168	26.07
Harpsichord Solo	Solo compositions for harpsichord	84	17419	43.93
Woodwind Ensemble	Ensembles consisting only of woodwind instruments	63	3784	10.05
Other Wind Ensemble	All kinds of wind ensembles beyond the woodwind family	51	3206	8.06

translations. Therefore, a massive dataset of image-audio pairs becomes a crucial enabler of our training strategy.

While there has been numerous datasets aligning multimodal music data using manual annotation or automatic audio-score alignment on either symbolic music scores [40]–[42] or sheet images [43], [44], they aimed to annotate the alignment densely, such as at the measure-level. However, if sparse alignment checkpoints—such as system- or slide-level transitions—are deemed sufficient, a far greater amount of data becomes accessible (Table II).

In particular, YouTube hosts an abundance of public *score-following* videos for Western classical music, where uploaders manually divide sheet music into slides, each typically containing two or three systems, and precisely synchronize slide transitions with the performance audio. These videos are primarily intended to help viewers follow along with the score while listening and are widely used for music education and enjoyment. As illustrated in Fig. 5, such videos inherently offer weak but musically meaningful alignment between score images and audio, providing a valuable and scalable resource for multimodal music translation.

A. Data Collection

We collect 12,217 score-following videos featuring various categories of western classical music. Each image-audio pair in a video forms a single data sample. Table I shows the statistics of the YouTube Score Video dataset.

We focus on Western classical music primarily due to data accessibility. In this tradition, the musical score is a standard published artifact that precedes performance, ensuring that high-quality sheet music is widely available. In contrast, for many other genres where music is distributed as audio, matching scores are often non-existent or manually transcribed retrospectively. Thus, classical music offers the most practical domain for collecting large-scale paired data.

While this data collection approach does not provide ground-truth symbolic representations, the resulting dataset contains a substantial amount of high-quality, in-the-wild sheet music and performance audio; to the best of our knowledge, it is the largest available collection of its kind.[1] Here we describe the various measures we take to extract and process the data samples.

B. Data Extraction and Processing

Slide Segmentation. We devise a system for identifying the timing points of slide transitions in the collected score-following videos. The transitions manifest in two distinct ways: instantaneous cuts between pages, and animated transitions such as crossfades or wipes. To handle this, we develop a rule-based segmentation algorithm that accommodates both cases while maintaining temporal accuracy. This process extracts the individual slides of score images along with their corresponding audio slices, creating the foundational paired segments for the dataset. Details are elaborated in Appendix A-C.

System Cropping and Resizing. The collected videos often feature letterboxes or pillarboxes, diverse aspect ratios, and inconsistent margins around contents; also, each slide contains an arbitrary number of musical systems alongside non-musical elements such as titles. To handle these irregularities, each musical system in a slide must be cropped and resized in a consistent format. Therefore, we label new annotations to fine-tune YOLOv8-based [45] models for two different tasks: system-wise bounding box regression, and staff line detection, as depicted in Fig. 6. We use the models to produce regularized crops of each system, a process which we discuss further in Appendix A-D.

C. Data Filtering

We filter the dataset in two levels: in the corpus-level, based on metadata; and in the sample-level, focusing on individual data quality.

Corpus-level Filtering. We employ the large language model Claude Sonnet 3.5 [46] to extract structured metadata from video titles, such as composer information, instrumentation, composition year, and category. Based on the collected

[1]The metadata for the dataset including links for each video, along with the codebase for preprocessing will be publicly released upon acceptance of the paper.

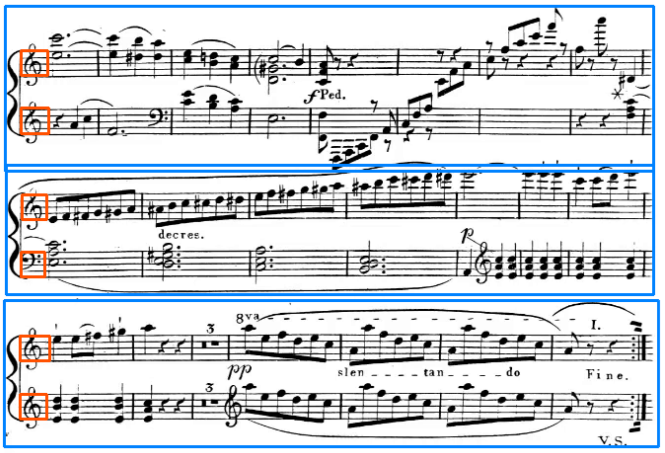

Fig. 6. Illustration of the music system detection pipeline using the fine-tuned YOLOv8 models. Detected systems are notated with blue boxes, and staff lines with red boxes. Note that the red boxes detect the staff height near clefs, not the clefs themselves.

metadata, we filter out orchestral compositions, pieces including singing voices, and some larger forms of chamber music. This is to rule out pieces that include a large number of staves per system or lyric texts, which would add excessive complexity to image tokenization. Additionally, we exclude pieces newer than the year 2000, as contemporary classical music often feature unconventional or highly complex notations. We further refine our dataset by removing handwritten or annotated scores. Details are elaborated in Appendix A-A and Appendix A-B.

Sample-Level Filtering. We address more granular aspects of data quality with algorithmic filtering. First, we implement pixel intensity-based metrics to identify and remove poor-quality scans and visual manipulations added by uploaders; second, we discard overly long or short audio samples, which might indicate errors in slide segmentation; finally, we filter cropped images based on their size and overlap with other bounding boxes to handle potential errors in system cropping. Details are elaborated in Appendix A-E.

For model training, we only use segments that are shorter than 20 seconds in audio and smaller than 256,000 total pixels in image. After filtering, we obtain 433,920 image-audio pairs from 12,217 videos, which amounts to 1,341 hours in total. This includes approximately 10,000 unique pieces by more than 2,000 composers.

VI. Dataset Compilation

The proposed multimodal, multi-task learning strategy requires a substantial amount of data in four distinct forms of music representation. Since existing datasets rarely cover all of the necessary modalities, we combine various datasets of different compositions to ensure adequate data in all four.

In addition to the YouTube Score Video dataset, we collect and process the following public datasets:

GrandStaff Dataset [47] (classical piano solo): a collection of synthetic pianoform **score images** and **notation**. It consists of solo piano repertoires from six composers: Beethoven, Scarlatti. The original 7,661 samples were augmented with six key augmentations, yielding 53,882 samples. Additionally, visual distortions were applied to generate paired distorted versions of the scores, resulting in 107,764 score-images in total. Every score is segmented to the system-level.

TABLE II
Data Distribution of Combined Datasets with Aligned Modalities.

Subset	Modalities				N	H
	Img	MXL	MIDI	Aud		
YTSV	√	-	-	√	433,920	1,341
GrandStaff	√	√	-	-	7,661	*23
OLiMPiC	√	√	-	-	17,945	*47
MusicNet$_{EM}$	-	-	△	√	330	33
MAESTRO	-	-	√	√	1,276	199
SLakh	-	-	√	√	2,100	145
BPSD	√	√	△	√	32	14

OLiMPiC Dataset [23] (piano accompaniment of classical art songs): contains scanned and synthetic pianoform **score images** and **notation**. Only synthetic images are included in the training set. The notation was sourced from the Open Score Lieder Corpus [48], which is crowd-sourced transcriptions of 19th-century piano-accompanied art songs, and only the piano staves are cropped from score images. The scores are segmented at the system level, resulting in a total of 17,945 samples.

MusicNet [42] (classical chamber music): dataset of **audio** and corresponding time-aligned **MIDI** annotations. We use the MIDI labels from MusicNet$_{EM}$ [21] that provide better alignment to the audio.

MAESTRO (classical piano solo) [3] : a collection of **audio** and **MIDI** data from piano performances. MIDI data is captured with Disklaviers (computer-controlled pianos), thus providing accurate timing precision.

SLakh [49] (pop music): a dataset of synthesized multi-track **audio** for the **MIDI** data in the Lakh dataset [50]. We only use the mixed audio and discard individual stems.

Beethoven Piano Sonata Dataset (BPSD) [51] (classical piano solo): **score images**, **notation**, multi-version **audio**, and their corresponding time-aligned **MIDI** of Beethoven's piano sonatas.

As BPSD does not provide system-level alignment between score images and other modalities, we manually annotate and align our own test subset, selecting 9 pieces that belong to the MAESTRO test set.

Table II shows the composition of the combined datasets used in our experiments, where △ denotes MIDI data made from audio-aligned scores, N denotes the number of aligned entries, and H denotes audio duration in hours. The symbolic notation datasets notated with * do not have audio files to report the duration. Therefore, for the OLiMPiC dataset, we computed the total duration by summing the audio lengths of every piece contained in the dataset, as provided by the official Open Score Lieder Corpus account on MuseScore [52]. Likewise, for the GrandStaff dataset, we obtained the total playing time by summing the MIDI durations of all pieces that constitute the dataset in KernScores [53]. The collected datasets, combined with the YouTube Score Video Dataset, forms the final assemblage used for model training.

VII. Experiments

A. Modal Directions

Although it is theoretically possible to train a single model that covers all modal translations, our preliminary experiments showed many practical difficulties. Instead, we train two different models in each of two directions described in Section II: *Image-to-Audio* (**I2A**) and *Audio-to-Image* (**A2I**). Each model is trained only for modal translations in its respective direction; for example, the I2A model is trained for image-to-audio, image-to-notation, and MIDI-to-audio tasks. Thus, I2A direction model has the input modality $\mathcal{X} \in \{\mathcal{I}, \mathcal{M}\}$ and the output modality $\mathcal{Y} \in \{\mathcal{N}, \mathcal{A}\}$, while A2I direction model has the opposite.

For the I2A model, we only focus on piano music instead of the entirety of chamber music, as modeling audio token sequences for various timbres proves challenging with the current dataset size. Also, the current OMR (image-to-LMX) dataset only covers piano repertoire, making it difficult to measure the effect of incorporating non-piano music in the other tasks; therefore, for image-to-audio translation, we use a subset of the YTSV dataset consisting of piano music (*YTSV-P*, 252k segments and 815 hours of audio), and only the MAESTRO dataset for MIDI-to-audio. For the A2I model, we instead use the entirety of the YTSV dataset.

B. Implementation

Our Transformer model uses 12 encoder layers and 12 decoder layers, with a model dimension of 1024, feed-forward hidden size of 4096 (with GELU activation) and 16 attention heads per layer, which are the same for both the I2A and A2I model variants. We initialize the token embedding matrix for the image and audio code tokens using the learned codebook embeddings from the RQVAE and DAC representation models, respectively, and initialize all other parameters randomly. We train each model using the AdamW optimizer [54] with an initial learning rate of 1×10^{-4}. The learning rate is scheduled to decay to 1×10^{-5} following a cosine decay schedule, after a linear warm-up of 2,000 steps. Training is run for 600,000 updates, with a total batch size of 24 sequence pairs, on 2× NVIDIA H100 SXM GPUs. For OMR, AMT and MIDI-to-audio tasks, the multi-task models were fine-tuned for the given task for 50k steps with a learning rate of $1\mathrm{e}{-5}$.

To handle the varying lengths of music pieces, we apply sequence-length truncation during training. For the I2A-direction model, we randomly slice each training sample so that audio clips are at most 20 seconds long (1723 tokens), and MIDI sequences are at most 1000 tokens long. For the A2I model, we use shorter truncations (10-second audio segments, up to 1000 MIDI tokens) to account for the more difficult audio-input tasks.

We also employ a curriculum learning strategy to gradually introduce the different tasks. Specifically, when training the I2A model, we begin with only the image-to-notation (OMR) samples in the batch mixture; after 15k training steps, we start including MIDI-to-audio synthesis; and after 50k steps, we begin adding the direct image-to-audio samples. Likewise, for the A2I model, we start with only audio-to-MIDI transcription, then add notation-to-image (score rendering) from 40k steps onward, and finally mix in audio-to-image after 70k steps. In this way, the model first learns easier or more data-rich subtasks before gradually facing the full complexity of the outer-modality translation tasks. Throughout training, we apply weighted sampling of the task datasets to balance their contributions, ensuring that no single task with a large dataset dominates the learning schedule. This training setup turns out to be crucial for stabilizing joint training and achieving good performance across all translation tasks.

C. Data Split and Test Sets

As most of our datasets focus on Western classical music repertoire, there exists a clear overlap of pieces featured between datasets; we carefully modify the assignments to avoid any overlap between splits. Among the YTSV dataset, pieces included in the test set of MAESTRO or MusicNet were assigned to the test set. We also make sure that multiple versions of a piece in YTSV are assigned to the same split.

The Beethoven Piano Sonata Dataset (BPSD) is the only subset in our collection that contains all four modalities, and we further preprocess a portion of it to create a test set for all translation tasks. Among the 32 sonatas in BPSD, we select 9 that are included in MAESTRO test split, and also exclude them from the training set of the GrandStaff dataset. After isolating the individual systems from the scanned images, we align each system with the corresponding sections of MusicXML, MIDI, and audio.

To evaluate I2A results on scanned scores of various quality, we manually select a subset of 11 pieces by 7 composers from the classical era to the contemporary (named *YTSV-T11*). We also manually check for duplications in the rest of YTSV. To conduct the human listening test, we manually select 12 musical systems from BPSD (named *BPSD-T12)*.

D. Evaluation Metrics

For image-to-audio generation, we evaluate the generated audio at the MIDI level using the following approach. One of the methods for assessing fidelity of performance synthesis is comparing the input condition and output by transcribing output audio to MIDI [4]. Given the reference and corresponding generated audio, we first apply the Onsets and Frames [15] model for piano transcription to obtain transcribed MIDI data. As the image-to-audio generation makes arbitrary choice of tempo expression, we use note onset, offset, and pitch information from the transcribed MIDIs to produce a time alignment between the reference and generated audio using dynamic time warping (DTW), following the approach from [21]. The alignment operates in one dimension by applying temporal warping across the entire piano roll along the time axis. This approach allows us to handle variations in tempo and timing while preserving the pitch information, enabling more accurate computation of onset F_1 scores by pairing corresponding events despite temporal differences. Fig. 7 illustrates this alignment process.

Finally, we compute the note onset F_1 score of the time-aligned reference and generated MIDI representations using

TABLE III
TRANSFORMER MODEL CONFIGURATIONS AND TASK INTRODUCTION STEPS. THE UPPER BLOCK (FIRST FIVE ROWS) SHOWS MODELS IN THE *Image-to-Audio* DIRECTION, AND THE LOWER BLOCK (LAST FOUR ROWS) MODELS IN THE *Audio-to-Image* DIRECTION. IN THE THREE "TASK INTRODUCTION" COLUMNS, PRE-SLASH TASKS ARE FOR I2A, AND THE POST-SLASH TASKS FOR A2I. NUMBERS INDICATE THE TRAINING STEP AT WHICH EACH TASK DATASET IS FIRST SAMPLED (0 = FROM THE START, "–" = NOT USED). M2A DENOTES MIDI-TO-AUDIO SYNTHESIS AND L2I DENOTES LMX-TO-IMAGE (SCORE RENDERING).

Models	Model Size				Task Introduction (trainging step)		
	Dimension	Enc/Dec Layers	Heads	Sub-Dec Heads	OMR/AMT	M2A/L2I	I2A/A2I
OMR Only	512	12	8	8	0	–	–
Image-to-Audio Only	768	12	10	10	–	–	0
MIDI-to-Audio Only	512	4	8	8	–	0	–
OMR + Image-to-Audio	1024	12	16	8	0	15,000	–
OMR + Image-to-Audio + MIDI-to-Audio	1024	12	16	8	0	15,000	50,000
AMT Only	768	12	12	8	0	–	–
Audio-to-Image Only	768	12	10	10	–	–	0
AMT + Audio-to-Image	1024	12	16	8	0	40,000	–
AMT + Audio-to-Image + LMX-to-Image	1024	12	16	8	0	40,000	70,000

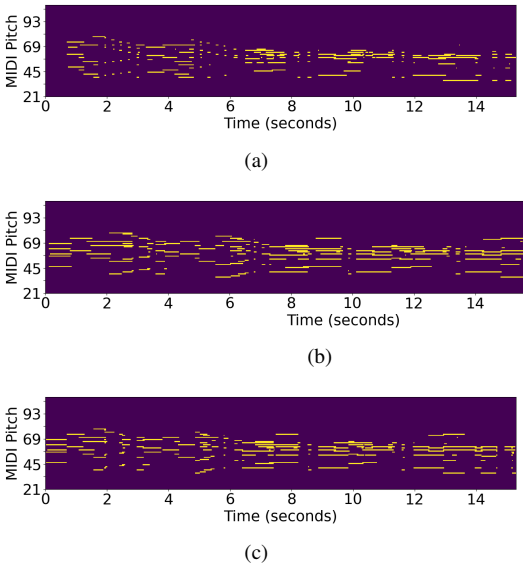

Fig. 7. Piano roll visualizations of MIDI representations from (a): Ground Truth, (b): Transcription from our Image-to-Audio generation model output, and (c): DTW-alignment of (b) on (a) for evaluation. The y-axis represents MIDI pitch values, spanning the full piano range from 21 to 108, encompassing all 88 pitch bins.

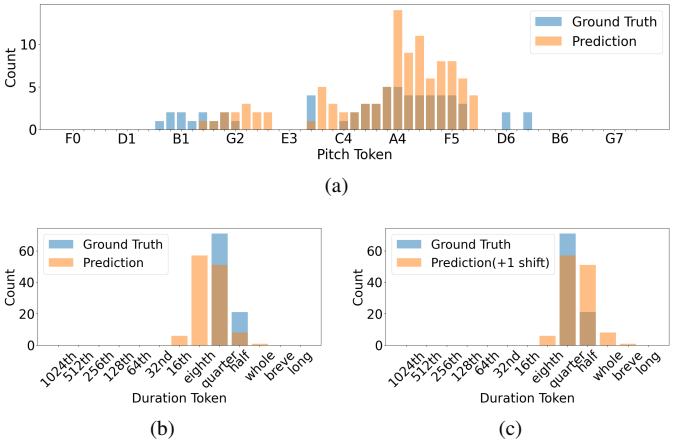

Fig. 8. Token distribution histograms used for Earth Mover's Distance (EMD) calculation. (a): Pitch token distribution histogram, (b): Duration token distribution histogram (no shift), (c): Duration token distribution histogram with +1 shift applied to prediction.

the transcription evaluation metrics from `mir_eval`. To account for potential DTW timing misalignments, we evaluate our results using three onset tolerance thresholds (50ms, 100ms, 200ms). We also report the Frechet Audio Distance (FAD) [55] between the generated results and the test set ground truths using the LAION-CLAP-audio embedding [56].

In addition to these objective metrics, we conduct a human listening test using 5-point mean opinion scores (MOS) to assess perceptual quality. For each test sample, participants are presented with a score image corresponding to a single system of piano music. Alongside the image, they are given the ground truth audio and the audio outputs generated by our proposed model and baseline systems. Participants are asked to rate each audio sample on a 5-point mean opinion score (MOS) scale (1-Worst, 5-Best) across the following three criteria:

- **Note Faithfulness (NF):** How accurately the audio reflects the musical notes written in the score (pitch, onset/offset, and notated content).
- **Performance Naturalness (PN):** How natural and musically plausible the rendition sounds as a performance (e.g., timing and dynamics).
- **Audio Quality (AQ):** The fidelity of the sound itself, disregarding musical elements (e.g., timbre, noise, and artifacts).

All raters evaluate an identical set of trials, *BPSD-T12*. To mitigate bias, the test is conducted in a single blind fashion, where raters are unaware of the origins of the audio clips (i.e., ground truth or a specific model). Furthermore, both the order of the trials and the order of the audio clips within each trial are randomized for each rater. We recruited 23 participants who regularly play classical music, and informed consent was obtained from all of them prior to evaluation.

For MIDI-to-audio synthesis we adopt an analogous protocol with image-to-audio evaluation, but omit the dynamic time-warping step: the reference MIDI is already precisely aligned to the generated audio, so we directly compute the note-onset F_1 score between the reference MIDI and the MIDI transcribed from the synthesis, and report the corresponding FAD under the same LAION-CLAP embedding.

For audio-to-image generation evaluation, we apply our reproduction of the Zeus [23] OMR model to extract LMX tokens from the audio-conditioned score image generations. We then compute the Earth Mover's Distance (EMD) [57] metric between histograms of the extracted LMX tokens

and the ground truth LMX. Specifically, we construct token frequency distributions where each bin position represents a unique token and bin height represents its frequency. To provide more granular analysis of our model's performance in capturing different musical aspects, we calculate EMD separately for pitch tokens (e.g., `C4`, `G3`, with 53 distinct pitch values) and duration tokens (e.g., `half`, `quarter`, `eighth`, with 13 distinct duration values). For duration tokens, which follow a 2:1 ratio between adjacent values (e.g., a half note is twice as long as a quarter note), we account for potential meter interpretation differences in our A2I model's output (such as single 4/4 measure interpreted as two measures with doubled note duration values). We address this by computing EMD with position shifts of -1, 0, and 1 on the predicted duration distribution and selecting the minimum EMD value among these shifts. As illustrated in Fig. 8, this shifting process can significantly improve the alignment between predicted and ground truth distributions, leading to a more accurate evaluation of the model's performance.

For OMR evaluation, we use the Symbol Error Rate (SER) metric on LMX token sequences, following [23]. For AMT evaluation, we use Note-F_1 score implemented in the `mir_eval` [58] library. As our main focus is western classical music, we exclude results for SLakh. For MusicNet, we use the string and wind test split following previous works [19], [21], [59].

E. Experimental Setup and Baselines

To validate the effectiveness of our unified, multi-task approach, we conduct a series of ablation studies by progressively adding more tasks to the training mixture. Furthermore, we evaluate our model's performance by comparing it against several external baseline models and conducting an internal comparison of different methodologies. The external baselines were chosen to represent the state-of-the-art in specialized, single-task domains.

For the novel image-to-audio (I2A) generation task, we perform two key comparisons. First, to demonstrate the effectiveness of our proposed *end-to-end* approach, we compare its performance directly against a *multi-stage procedural cascade* that utilizes our own unified model. In this setup, the model first performs OMR (image to MusicXML), which is then converted to MIDI, and finally the same model synthesizes the audio. This internal comparison is designed to highlight the differences between the two methodologies. Second, as a multi-stage baseline for external comparison, we constructed a pipeline composed of specialized, state-of-the-art models: Zeus [23] for OMR, VirtuosoNet (VNet) [8] for performance modeling, and Music-Spectrogram Diffusion (MSD) [4] for audio synthesis.

For the MIDI-to-audio synthesis task, we use the MSD [4] model as a baseline for comparison, as it represents a state-of-the-art method for multi-track MIDI-conditioned audio generation.

For the Optical Music Recognition (OMR) task, we compare our model against a reproduction of Zeus [23], the state-of-the-art model for pianoform OMR. This model was chosen as a baseline because it is a state-of-the-art system for pianoform OMR that, similar to our approach, operates on single musical systems. Since the original Zeus paper did not report results on the combined GrandStaff and OLiMPiC datasets, we trained our own version of the model on the exact same data split used for our experiments to ensure a fair and direct comparison.

Our Automatic Music Transcription (AMT) results are compared with previously published state-of-the-art models. We chose Maman *et al.* [21] and Chang *et al.* (`YPTF.MoE+M`) [4] because they are capable of multi-track MIDI transcription for a variety of instruments, not limited to solo piano. Specifically, Maman *et al.* which used additional 73 hours of unaligned score and audio for training. Chang *et al.*'s model represents a highly specialized and optimized architecture for AMT.

VIII. Results

A. Image-to-Audio Generation

Generating music audio directly from a given score image has never been explored in an end-to-end manner. Our result in Table IV shows that our I2A model trained with multiple tasks generates musically-coherent performance audio that corresponds to the input score image, showing high F_1 scores. To demonstrate the performance bottleneck from audio tokenization, we also measure the metric for DAC-reconstructed audio of reference performance, which can be regarded as the upper limit.

As mentioned earlier, the model only trained for image-to-audio task, *YTSV-P (I2A Only)* could not generate audio that is coherent to input score image as shown by low transcription result nor fluent audio as shown by high FAD. On the other hand, the model trained along with the OMR task (*OMR + I2A*) showed significant improvement in both transcription score and FAD, showing clear evidence that the model read the notes from input score image to generate corresponding audio. Adding MIDI-to-audio (M2A) task (*OMR + I2A + M2A*) improved note reconstruction accuracy at YTSV test set. The model combined with the MIDI-to-audio task showed an improved FAD score, which indicates that the model learned to model more natural sound. Overall, the results show that encompassing different modal translation tasks improves performance.

When comparing the *Multi-stage* approach to the *Direct I2A* method on the same model (*OMR + I2A + M2A*), the F_1 scores were better, but the corresponding FAD value was worse. In practice, this translates to more accurate note timing and pitch on paper, yet a noticeably lower perceptual quality—often manifesting as dissonant or "off-sounding" passages. The bottleneck mainly comes from OMR errors. We found that errors in the OMR stage often result in perceptually obvious flaws, such as strong dissonance. This stems from a fundamental limitation of current OMR models: they are trained to optimize symbol-level accuracy without any awareness of the downstream auditory impact. Visually subtle errors, such as misread accidentals or notes misclassified by a semitone, can lead to musically severe consequences in the audio output.

The results from our human evaluation, conducted with 23 raters, corroborate our objective findings and highlight the

TABLE IV
IMAGE-TO-AUDIO ACCURACY REPORTED AS ONSET F_1 (↑), FAD (↓) AND MOS ↑. NOTE THAT THE TWO ROWS MARKED *OMR + I2A + M2A* SHARE A **SINGLE** MODEL JOINTLY TRAINED ON THOSE THREE TASKS. *Direct I2A* FEEDS A SCORE IMAGE TO THE MODEL AND PREDICTS AUDIO IN ONE SHOT, WHEREAS *Multi-stage* FIRST APPLIES THE MODEL FOR OMR (IMAGE → MUSICXML), CONVERTS THAT MUSICXML TO MIDI WITH A SIMPLE RULE-BASED CONVERSION WITH FIXED TEMPO AND MIDI, AND THEN REUSES THE SAME MODEL FOR M2A (MIDI → AUDIO). SINCE YTSV-T11 CONSISTS OF MULTI-SYSTEM IMAGES (I.E. HAVING TWO OR MORE SYSTEMS PER IMAGE), WE DO NOT REPORT MULTI-STAGE RESULTS AS THE MODEL WAS NOT TRAINED FOR MULTI-SYSTEM OMR.

Method	Metric	F_1 Score ↑						FAD ↓		MOS ↑		
	Dataset	BPSD			YTSV-T11			BPSD	YTSV-T11	BPSD-T12		
	Onset Tolerance / Criteria	50ms	100ms	200ms	50ms	100ms	200ms	—	—	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

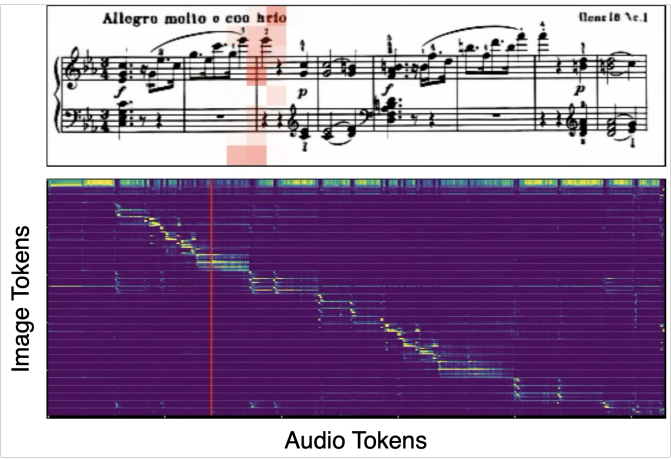

Fig. 9. Attention patterns from a selected transformer head. Top: Red heatmap shows attention weights across RVQ token regions in the image. Bottom: Token-level attention matrix with audio tokens (x-axis) versus image tokens (y-axis), revealing cross-modal relationships.

perceptual superiority of the end-to-end design. The MOS listening test results establish our *Direct I2A* model as the clear preference among human evaluators, outperforming other methods across all criteria. Crucially, the ratings validate the perceptual weakness of the *Multi-stage* approach suggested by the FAD metric. Despite achieving higher F_1 scores, the cascaded system was perceived as significantly less faithful and natural, supporting the hypothesis that it propagates musically critical OMR errors. In contrast, the end-to-end model learns to avoid such perceptually jarring artifacts, producing a more musically coherent output.

To validate that the model actually attends to the corresponding parts of score image to generate performance audio, we visualize the attention map of the model as shown in Fig. 9. The result clearly shows that the model follows the score to translate it to performance audio. To evaluate the quality of the generated results, we strongly encourage readers to refer to the actual audio and video examples provided on our demo page[2].

[2]https://sakem.in/u-must/
Results of another I2A model trained on YTSV chamber music pieces is also included in the bottom of the demo page.

TABLE V
AUDIO-TO-IMAGE GENERATION ACCURACY IN EMD ON BPSD.

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

A qualitative analysis of the generation results highlights two primary sources of error. First, due to the inherent loose alignment in the source videos used to construct the YTSV dataset, training pairs sometimes contain audio segments that are longer or shorter than the depicted score. Consequently, the model frequently generates additional, improvisation-like music after the audio corresponding to the score concludes, which introduces false positives and negatively affects the F_1 score. Second, the model struggles with acoustic clarity in passages featuring dense musical textures, where individual notes are not always distinctly synthesized. As a result, we observe lower F_1 scores in fast-paced clips characterized by a high note density.

B. Audio-to-Image Generation

Converting audio recordings into sheet music requires navigating the complex transformation from acoustic signals to structured visual notation. Our results in Table V highlight the progress of our A2I model in this challenging domain. To contextualize our evaluation metrics, we include measurements for images reconstructed using the RQVAE, ground truth images processed through our OMR system, and a ground truth random pairing baseline–where we compute the EMD between one ground truth sample and the remaining ground truth samples–as reference points. Interestingly, the ground truth variability baseline indicates better performance than the A2I model trained without additional tasks, underscoring the difficulty of the A2I task without further guidance. The integration of auxiliary tasks proves beneficial for A2I performance; our model, trained with additional AMT and LMX-

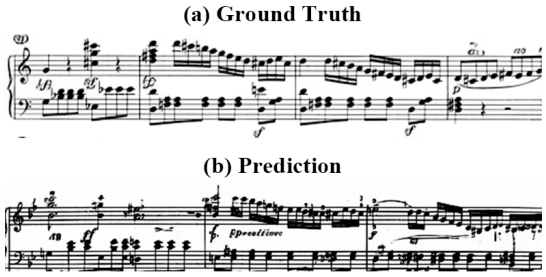

Fig. 10. One example from audio-to-image translation.

to-image tasks, shows improved EMD metrics compared to the baseline trained solely on A2I, suggesting that explicitly modeling intermediate symbolic representations enhances the model's ability to generate appropriate musical notation.

The consistently higher pitch EMD values compared to duration EMD values observed in the table reflect the structural difference in token space sizes rather than indicating lower pitch accuracy; with 53 distinct pitch tokens versus only 13 duration tokens, EMD naturally yields larger values when measuring transportation distances across the wider pitch space.

Although the generated result does not yet fully satisfy human-readable and performable quality standards, it does demonstrate that the A2I model partially reflects the notes played in the audio, as shown in Fig. 10. The asymmetry in generation quality between I2A and A2I likely stems from the intrinsic difficulty of generating systematic notation directly from performance audio—a challenge compounded by the need for complex quantization and metric alignment in an end-to-end setting [60].

Furthermore, generating score images introduces unique hurdles; unlike natural images, musical scores require strict structural precision where minor visual deviations can drastically alter the musical meaning. We also note a disparity in data augmentation strategies. While comprehensive geometric augmentations (e.g., distortion, warping) were essential for image-input tasks to bridge the domain gap between software-rendered and scanned scores, applying such perturbations to the generation targets is unfeasible. Doing so would artificially broaden the target distribution, introducing excessive variance that can destabilize the training of the generative model. Therefore, this creates a significant domain gap. The model is trained to generate pristine, software-rendered images from symbols, while simultaneously learning to generate irregular, scanned-style images from audio. This inconsistency in the target distributions likely reduced the overall synergy of multi-task training.

To mitigate these challenges, we believe a promising future direction lies in leveraging OMR to generate symbolic pseudo-MusicXML from our YTSV dataset. Incorporating this symbolic modality would serve a dual purpose: it acts as a synergistic auxiliary target to bypass the hurdles of direct pixel generation, while simultaneously addressing real-world utility by providing an editable, practical output format (Audio-to-Symbolic) alongside the visual score.

TABLE VI
MIDI-TO-AUDIO SYNTHESIS ACCURACY IN F_1 AND FAD ON BPSD.

Method	$F_1 \uparrow$			FAD $\downarrow$
	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 [4]	83.82	90.63	92.28	0.229

C. MIDI-to-Audio Synthesis

Table VI summarizes MIDI-to-Audio (M2A) performance on the Beethoven Piano Sonata Dataset. A model trained solely on M2A achieves onset F_1 scores of 26.61, 64.86, and 88.20 in 50ms, 100ms, and 200ms tolerance windows, respectively, with a Fréchet Audio Distance (FAD) of 0.201. Augmenting training with OMR and Image-to-Audio tasks yields pronounced gains at stricter tolerances: the 50ms F_1 rises to 45.91 (+19.3 pp) and the 100 ms score to 69.54 (+4.7 pp), while the 200ms result remains comparable (84.55). Concomitantly, FAD decreases by 42 % to 0.116, indicating markedly improved perceptual quality. These findings confirm that incorporating complementary score-conditioned tasks sharpens temporal precision and enhances audio realism in M2A synthesis.

Our model demonstrates a lower F_1 score compared to the MSD baseline [4]. We hypothesize that the autoregressive generation of audio tokens is inherently less robust in maintaining strict temporal alignment compared to diffusion-based architectures. However, our model achieves a superior FAD score. We believe this is due to the difference in training data; our model was trained on only the MAESTRO dataset for the MIDI-to-audio task, whereas MSD was trained on a broader range of datasets, many of which are synthetic. Consequently, the audio distribution of our model, trained exclusively on real piano performances, is likely closer to that of the BPSD test set, resulting in a better FAD score.

D. OMR and AMT

To validate the effectiveness of incorporating image-to-audio and audio-to-image for solving other tasks, we also train models using only OMR datasets (GrandStaff and Olimpic) or AMT datasets (MAESTRO, MusicNet, and SLakh). As the size of the datasets differs greatly from YTSV, we experiment with various model sizes and select the one with the best performance on the validation set. The selected baseline OMR-only model consists of an encoder and a decoder, each with 12 layers, 512 dimensions, and 8 attention heads, and the AMT-only model consists of the same number of layers but with 768 dimensions and 12 attention heads.

Table VII provides OMR evaluation results on OLiMPiC and BPSD-test. The results show that performance increases as more modal translation tasks were added, and our unified approach achieves the best results. The addition of MIDI-to-audio translation, a task that has no modal overlap with OMR, also improves the performance of OMR. We hypothesize that this is because the model is trained to follow note-level conditions in MIDI-to-audio, which guides the model to better handle similar information in score image inputs.

TABLE VII
OMR RESULTS IN SER. LOWER IS BETTER.

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

TABLE VIII
AMT RESULTS IN NOTE ONSET F_1 SCORE FOR TEST SET. HIGHER IS BETTER.

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.* [21]	80.0	87.5	89.7
Chang *et al.* [19]	91.32	83.46	96.98

Table VIII presents the results of AMT in terms of Note-F_1 scores. Unlike in OMR, incorporating audio-to-image translation does not lead to noticeable improvements. We attribute this to the already extensive training data available for AMT such as MAESTRO, which provides 200 hours of piano audio with precisely aligned MIDI annotations. Our model performs better than Maman *et al.* [21] for string instruments even without additional training, while performing worse in woodwind. We believe the difference comes from dataset distribution. While the dataset used by Maman *et al.* included 26 hours of orchestral recordings, our dataset did not, thus heavily lacking woodwind sounds; on the other hand, the higher proportion of string instrument samples in our dataset may have provided our representation model with a learning advantage in capturing the distinct timbre characteristics. Furthermore, the `YPTF.MoE+M(noPS)` model from Chang *et al.* [19] utilizes extensive augmentations like mixup and employs a domain-specific encoder architecture tailored for the AMT task. In contrast, our approach pre-quantizes audio into discrete RVQ tokens to facilitate the unified framework. This design choice, while crucial for our multimodal translation goal, prevents the application of on-the-fly augmentations, which likely accounts for the performance gap compared to the highly specialized YMT3+ model. Finally, the minimal impact of the Audio-to-Image task on AMT performance may also stem from the suboptimal quality of our A2I generation; unlike the robust I2A results, the A2I task did not achieve sufficient maturity to provide meaningful synergistic feedback for transcription.

IX. CONCLUSION

In this work, we presented a unified methodology for music modal translation based on a common tokenization strategy capable of musically-coherent score image-to-audio generation, along with state-of-the-art performance in optical music recognition. We believe that this result highlights the potential of integrating various modalities for a more holistic understanding and processing of music information.

We also presented the YouTube Score Video (YTSV) dataset, a large-scale collection of score image and performance audio pairs. In particular, our experiment results with YTSV emphasize the importance of scanned music scores as a valuable source of musical data, an area that has received relatively less attention in music information retrieval compared to the audio and symbolic domains. Given the vast number of publicly available scanned scores in the International Music Score Library Project (IMSLP), we expect that incorporating these resources could have a significant impact on music understanding and generation.

However, there are clear limitations for the current study. First, the improvement in AMT accuracy is modest relative to the substantial gains observed for OMR and I2A translation. Second, the quality of the A2I outputs remains insufficient for practical use: rendered pages frequently exhibit overcrowded staves, misplaced clefs, and other notational artifacts. Third, our current framework utilizes distinct models for the I2A and A2I directions, sharing an identical architecture. While this allows for task-specific optimization, consolidating these into a single omnipotent model remains a compelling avenue for future research. While the current framework employed the encoder-decoder structure, a decoder-only model could be considered as a powerful tool for unifying all modal directions into a single model.

Looking forward, our YTSV dataset opens several promising research directions. We plan to explore self-supervised learning for more precise image-audio alignment, develop score-conditioned audio tasks such as expressive performance modeling, and investigate methods to extend sparse slide-level alignments to note-level supervision. Additionally, improving the underlying audio and image codecs remains an important goal to enhance both the visual quality of our A2I outputs and the acoustic fidelity of I2A generation.

REFERENCES

[1] E. Benetos, S. Dixon, Z. Duan, and S. Ewert, "Automatic music transcription: An overview," *IEEE Signal Processing Magazine*, vol. 36, no. 1, pp. 20–30, 2019.

[2] F. Jamshidi, G. Pike, A. Das, and R. Chapman, "Machine learning techniques in automatic music transcription: A systematic survey," 06 2024.

[3] 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.

[4] C. Hawthorne, I. Simon, A. Roberts, N. Zeghidour, J. Gardner, E. Manilow, and J. Engel, "Multi-instrument music synthesis with spectrogram diffusion," in *Proceedings of the 23rd International Society for Music Information Retrieval Conference (ISMIR 2022)*, 2022.

[5] J. Calvo-Zaragoza, J. Hajič, jr, and A. Pacha, "Understanding optical music recognition," *ACM Computing Surveys*, vol. 53, 05 2020.

[6] J. Calvo-Zaragoza, J. Martinez-Sevilla, C. Peñarrubia, and A. Ríos Vila, *Optical Music Recognition: Recent Advances, Current Challenges, and Future Directions*, pp. 94–104. 08 2023.

[7] E. Nakamura, E. Benetos, K. Yoshii, and S. Dixon, "Towards complete polyphonic music transcription: Integrating multi-pitch detection and rhythm quantization," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 101–105, IEEE, 2018.

[8] D. Jeong, T. Kwon, Y. Kim, and J. Nam, "Graph neural network for music score data and modeling expressive piano performance," in *International conference on machine learning*, pp. 3060–3070, PMLR, 2019.

[9] L. Liu, Q. Kong, V. Morfi, and E. Benetos, "Performance midi-to-score conversion by neural beat tracking," in *International Society for Music Information Retrieval Conference*, 2022.

[10] T. Beyer and A. Dai, "End-to-end piano performance-midi to score conversion with transformers," in *International Society for Music Information Retrieval Conference*, 2024.

[11] H.-W. Dong, C. Zhou, T. Berg-Kirkpatrick, and J. McAuley, "Deep performer: Score-to-audio music performance synthesis," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 951–955, IEEE, 2022.

[12] J. Tang, E. Cooper, X. Wang, J. Yamagishi, and G. Fazekas, "Towards an integrated approach for expressive piano performance synthesis from music scores," in *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2025.

[13] M. Kim, J.-w. Jung, H. Rha, S. Maiti, S. Arora, X. Chang, S. Watanabe, and Y. M. Ro, "Tmt: Tri-modal translation between speech, image, and text by processing different modalities as different languages," *arXiv preprint arXiv:2402.16021*, 2024.

[14] E. Benetos, S. Dixon, Z. Duan, and S. Ewert, "Automatic music transcription: An overview," *IEEE Signal Processing Magazine*, vol. 36, no. 1, pp. 20–30, 2019.

[15] C. Hawthorne, E. Elsen, J. Song, A. Roberts, I. Simon, C. Raffel, J. H. Engel, S. Oore, and D. Eck, "Onsets and frames: Dual-objective piano transcription," in *Proceedings of the 19th International Society for Music Information Retrieval Conference, ISMIR 2018, Paris, France, September 23-27, 2018* (E. Gómez, X. Hu, E. Humphrey, and E. Benetos, eds.), pp. 50–57, 2018.

[16] Q. Kong, B. Li, X. Song, Y. Wan, and Y. Wang, "High-resolution piano transcription with pedals by regressing onset and offset times," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3707–3717, 2021.

[17] C. Hawthorne, I. Simon, R. Swavely, E. Manilow, and J. H. Engel, "Sequence-to-sequence piano transcription with transformers," in *Proceedings of the 22nd International Society for Music Information Retrieval Conference, ISMIR 2021, Online, November 7-12, 2021* (J. H. Lee, A. Lerch, Z. Duan, J. Nam, P. Rao, P. van Kranenburg, and A. Srinivasamurthy, eds.), pp. 246–253, 2021.

[18] 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.

[19] 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.

[20] Z. Wang, D. Xu, G. Xia, and Y. Shan, "Audio-to-symbolic arrangement via cross-modal music representation learning," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 181–185, 2022.

[21] B. Maman and A. H. Bermano, "Unaligned supervision for automatic music transcription in the wild," in *Proceedings of the 39th International Conference on Machine Learning* (K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato, eds.), vol. 162 of *Proceedings of Machine Learning Research*, pp. 14918–14934, PMLR, 17–23 Jul 2022.

[22] A. Ríos-Vila, J. Calvo-Zaragoza, and T. Paquet, "Sheet music transformer: End-to-end optical music recognition beyond monophonic transcription," in *International Conference on Document Analysis and Recognition*, pp. 20–37, Springer, 2024.

[23] 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.

[24] 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.

[25] A. Ríos-Vila, J. Calvo-Zaragoza, D. Rizo, and T. Paquet, "Sheet music transformer++: End-to-end full-page optical music recognition for pianoform sheet music," *arXiv preprint arXiv:2405.12105*, 2024.

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

[27] Y. Lin, Z. Dai, and Q. Kong, "Musicscore: A dataset for music score modeling and generation," 2024.

[28] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," in *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA* (I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, and R. Garnett, eds.), pp. 6626–6637, 2017.

[29] E. Shatri, K. R. Palavala, and G. Fazekas, "Synthesising handwritten music with gans: A comprehensive evaluation of cyclewgan, progan, and DCGAN," in *IEEE International Conference on Big Data, BigData 2024, Washington, DC, USA, December 15-18, 2024* (W. Ding, C. Lu, F. Wang, L. Di, K. Wu, J. Huan, R. Nambiar, J. Li, F. Ilievski, R. Baeza-Yates, and X. Hu, eds.), pp. 3208–3217, IEEE, 2024.

[30] X. He, X. Han, L. Wei, L. Xie, and Q. Tian, "Mixpert: Mitigating multimodal learning conflicts with efficient mixture-of-vision-experts," 2025.

[31] H. Ye and D. Xu, "Taskexpert: Dynamically assembling multi-task representations with memorial mixture-of-experts," in *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pp. 21771–21780, IEEE, 2023.

[32] W. Yu, H. Xu, Z. Yuan, and J. Wu, "Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis," in *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pp. 10790–10797, AAAI Press, 2021.

[33] D. Yang, J. Tian, X. Tan, R. Huang, S. Liu, H. Guo, X. Chang, J. Shi, S. Zhao, J. Bian, Z. Zhao, X. Wu, and H. M. Meng, "UniAudio: Towards universal audio generation with large language models," in *Proceedings of the 41st International Conference on Machine Learning* (R. Salakhutdinov, Z. Kolter, K. Heller, A. Weller, N. Oliver, J. Scarlett, and F. Berkenkamp, eds.), vol. 235 of *Proceedings of Machine Learning Research*, pp. 56422–56447, PMLR, 21–27 Jul 2024.

[34] R. Valle, R. Badlani, Z. Kong, S. gil Lee, A. Goel, S. Kim, J. F. Santos, S. Dai, S. Gururani, A. Aljafari, A. H. Liu, K. J. Shih, R. Prenger, W. Ping, C.-H. H. Yang, and B. Catanzaro, "Fugatto 1: Foundational generative audio transformer opus 1," in *The Thirteenth International Conference on Learning Representations*, 2025.

[35] W. Zeng, J. Zhao, and Y. Wang, "Disentangling score content and performance style for joint piano rendering and transcription," 2025.

[36] C. Fan, W. Xiang, J. Tao, J. Yi, and Z. Lv, "Cross-modal knowledge distillation with multi-stage adaptive feature fusion for speech separation," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 33, pp. 935–948, 2025.

[37] 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.

[38] 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.

[39] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, vol. 30, pp. 6000–6010, 2017.

[40] C. Weiß, F. Zalkow, V. Arifi-Müller, M. Müller, H. V. Koops, A. Volk, and H. G. Grohganz, "Schubert winterreise dataset: A multimodal scenario for music analysis," *Journal on Computing and Cultural Heritage*, vol. 14, May 2021.

[41] C. Weiß, V. Arifi-Müller, M. Krause, F. Zalkow, S. Klauk, R. Kleinertz, and M. Müller, "Wagner ring dataset: A complex opera scenario for music processing and computational musicology," *Transactions of the International Society for Music Information Retrieval*, vol. 6, no. 1, 2023.

[42] J. Thickstun, Z. Harchaoui, and S. Kakade, "Learning features of music from scratch," in *International Conference on Learning Representations*, 2017.

[43] D. Yang, T. Tanprasert, T. Jenrungrot, M. Shan, and T. Tsai, "MIDI passage retrieval using cell phone pictures of sheet music," in *Proceedings of the 20th International Society for Music Information Retrieval Conference, ISMIR 2019, Delft, The Netherlands, November 4-8, 2019* (A. Flexer, G. Peeters, J. Urbano, and A. Volk, eds.), pp. 916–923, 2019.

[44] M. Feffer, C. Donahue, and Z. Lipton, "Assistive alignment of in-the-wild sheet music and performances," in *Late-Breaking/Demo of 23rd International Society for Music Information Retrieval Conference (ISMIR)*, 2022.

[45] 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.
[46] Anthropic, "Claude [large language model]," 2023.
[47] 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.
[48] M. Gotham and P. Jonas, "The openscore lieder corpus," in *Music Encoding Conference Proceedings*, Universidad de Alicante, 2021.
[49] E. Manilow, G. Wichern, P. Seetharaman, and J. Le Roux, "Cutting music source separation some Slakh: A dataset to study the impact of training data quality and quantity," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, IEEE, 2019.
[50] C. Raffel, *Learning-Based Methods for Comparing Sequences, with Applications to Audio-to-MIDI Alignment and Matching*. PhD thesis, COLUMBIA UNIVERSITY, 2016.
[51] J. Zeitler, C. Weiß, V. Arifi-Müller, and M. Müller, "Bpsd: A coherent multi-version dataset for analyzing the first movements of beethoven's piano sonatas," *Transactions of the International Society for Music Information Retrieval*, vol. 7, no. 1, 2024.
[52] M. R. H. Gotham and P. Jonas, "The OpenScore Lieder Corpus," in *Music Encoding Conference Proceedings 2021* (S. Münnich and D. Rizo, eds.), pp. 131–136, Humanities Commons, 2022.
[53] "KernScores — kern.humdrum.org." https://kern.humdrum.org/. [Accessed 09-05-2025].
[54] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *International Conference on Learning Representations*, 2019.
[55] A. Gui, H. Gamper, S. Braun, and D. Emmanouilidou, "Adapting frechet audio distance for generative music evaluation," in *Proc. IEEE ICASSP*, 2024.
[56] Y. Wu, K. Chen, T. Zhang, Y. Hui, T. Berg-Kirkpatrick, and S. Dubnov, "Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation," in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, 2023.
[57] Y. Rubner, C. Tomasi, and L. J. Guibas, "The earth mover's distance as a metric for image retrieval," *International Journal of Computer Vision*, vol. 40, pp. 99–121, Nov 2000.
[58] C. Raffel, B. McFee, E. J. Humphrey, J. Salamon, O. Nieto, D. Liang, D. P. Ellis, and C. C. Raffel, "Mir_eval: A transparent implementation of common mir metrics.," in *ISMIR*, vol. 10, p. 2014, 2014.
[59] K. W. Cheuk, D. Herremans, and L. Su, "Reconvat: A semi-supervised automatic music transcription framework for low-resource real-world data," in *Proceedings of the 29th ACM International Conference on Multimedia*, pp. 3918–3926, 2021.
[60] A. Galan-Cuenca, J. J. Valero-Mas, J. C. Martinez-Sevilla, A. Hidalgo-Centeno, A. Pertusa, and J. Calvo-Zaragoza, "Muscat: a multimodal music collection for automatic transcription of real recordings and image scores," in *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 583–591, 2024.

X. Biography Section

Jongmin Jung received both M.S. degree in Artificial Intelligence (2025) and B.A.S (Bachelor of Arts and Science) degree in Art & Technology (2022) from Sogang University in Seoul, South Korea. His research spans music information retrieval tasks, multimodal music representation, generative modeling, and deep-neural-network–driven audiovisual synthesis.

Dongmin Kim received his B.A.S. degree in Art & Technology from Sogang University, Seoul, South Korea in 2020, graduating Magna Cum Laude. He is currently pursuing an M.S. degree in Artificial Intelligence at Sogang University. His research interests include Music Information Retrieval (MIR) and Optical Music Recognition (OMR).

Sihun Lee received his B.A.S degree in Art & Technology from Sogang University in 2023, and is currently in his Masters program in artificial intelligence at Sogang University. His main research interest is practical, human-centric implementation of neural models in Music Information Retrieval (MIR).

Seola Cho received the B.A. in Creative Writing from Dongduk Women's University, Seoul, South Korea, in 2022. She is currently a researcher in FutureLab, and an intern in MALerLab, at Sogang University. Her research interests include automatic music transcription and symbolic music generation.

Hyungjoon Soh received the B.S. degree in Physics in 2011 and the Ph.D. degree in Physics in 2018, both from the Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Korea. From 2018 to 2021, he was a Research Scientist with Kakao Brain, Seongnam, Korea, where he worked on speech synthesis and large-scale language modeling. Since 2022, he has been a Post-Doctoral Researcher with the Department of Physics Education, Seoul National University, Seoul, Korea. His current research focuses on physics-inspired models for improving sequence-to-sequence representation and generation in speech, natural language, and symbolic music.

Irmak Bukey received her B.A. degree in Computer Science and Mathematics from Pomona College in Claremont, CA in 2023. She is currently a Ph.D. candidate at Carnegie Mellon University in Pittsburgh, PA at the Generative Creativity Lab. Her research mainly lies at the intersection of music and machine learning. Her research interests include music information retrieval, audio signal processing and alignment, and multimodal music representation and generation.

Chris Donahue received the Ph.D. degree from the University of California San Diego, La Jolla, CA, USA, where he was jointly advised by Miller Puckette(music) and Julian McAuley (CS). He was a Postdoctoral Scholar with the Computer Science Department, Stanford University, advised by Percy Liang. He is currently an Assistant Professor with Computer Science Department, Carnegie Mellon University, Pittsburgh, PA, USA, and a part-time research Scientist with Google DeepMind, London, U.K., working on the Magenta project. His research goal is to develop and responsibly deploy generative AI for music and creativity, thereby unlocking and augmenting human creative potential.

Dasaem Jeong is currently working as an Assistant Professor in the Department of Art & Technology at Sogang University in South Korea since 2021. He obtained his Ph.D. and M.S. degrees in culture technology, and B.S. in mechanical engineering from Korea Advanced Institute of Science and Technology (KAIST). His research primarily focuses on a diverse range of music information retrieval tasks with special interests on Western classical music and Korean traditional music.