DOI: 10.1111/cgf.70025
EUROGRAPHICS 2025 / A. Bousseau and A. Dai (Guest Editors)

COMPUTER GRAPHICS *forum*
Volume 44 (2025), Number 2

Multi-Modal Instrument Performances (MMIP): A Musical Database

T. Kyriakou[1,2] A. Aristidou[1,2] and P. Charalambous[2]

[1]University of Cyprus, Nicosia, Cyprus
[2]CYENS Centre of Excellence, Nicosia, Cyprus

Abstract

Musical instrument performances are multimodal creative art forms that integrate audiovisual elements, resulting from musicians' interactions with instruments through body movements, finger actions, and facial expressions. Digitizing such performances for archiving, streaming, analysis, or synthesis requires capturing every element that shapes the overall experience, which is crucial for preserving the performance's essence. In this work, following current trends in large-scale dataset development for deep learning analysis and generative models, we introduce the Multi-Modal Instrument Performances (MMIP) database (https://mmip.cs.ucy.ac.cy). This is the first dataset to incorporate synchronized high-quality 3D motion capture data for the body, fingers, facial expressions, and instruments, along with audio, multi-angle videos, and MIDI data. The database currently includes 3.5 hours of performances featuring three instruments: guitar, piano, and drums. Additionally, we discuss the challenges of acquiring these multi-modal data, detailing our approach to data collection, signal synchronization, annotation, and metadata management. Our data formats align with industry standards for ease of use, and we have developed an open-access online repository that offers a user-friendly environment for data exploration, supporting data organization, search capabilities, and custom visualization tools. Notable features include a MIDI-to-instrument animation project for visualizing the instruments and a script for playing back FBX files with synchronized audio in a web environment.

CCS Concepts
• ***Computing methodologies*** → ***Motion capture; Motion processing;*** *Mixed / augmented reality; Virtual reality; Machine learning;* • ***Applied computing*** → ***Performing arts; Digital libraries and archives;*** • ***Information systems*** → ***Information retrieval;***

1. Introduction

Datasets play a critical role in advancing the field of computer graphics and computer vision, providing the foundation for training, analysis, testing, and the development of new algorithms and technologies. Well-known datasets in human animation, such as AMASS [MGT*19], CMU [CMU03] and Human3.6M [IPOS14], have pushed the boundaries of motion analysis and synthesis, enabling the development of more advanced algorithms. High-quality datasets has also driven significant advancements in deep learning and generative models, leading to transformative applications such as text-to-image generation (e.g., DALL-E [RPG*21]), text-to-video synthesis (e.g., SORA [BPH*24]), and AI-powered applications like ChatGPT [Ope22].

A growing trend toward multi-modal databases has emerged in recent years, reflecting the increasing complexity of research challenges. These datasets integrate audio, vision, pose, and annotation, that are crucial for generative models. For example, EMHI [FDS*24] and the Nymeria [MYH*24] dataset combine multiple signal inputs, supporting advancements in areas such as body tracking and action recognition.

Despite the rapid growth in multi-modal datasets, a significant gap still remains in the digital capture of human musical performances. One of the primary reasons this area is underdeveloped is the complexity and difficulty involved in capturing and curating such data. Musical performances is a highly multi-modal art form, involving not only auditory elements but also a rich array of visual and physical signals capturing all of these can involve 3D motion capture of body movements, facial expressions, and precise finger positions, along with high-quality audio, multi-angle video, and symbolic representations like MIDI (Musical Instrument Digital Interface - more information about MIDI can be found in Section 4.3). Synchronizing and aligning these diverse data streams is a major technical challenge, requiring specialized equipment and methodologies. The variety of instruments, playing techniques, and performance styles adds further complexity to the creation of a

comprehensive dataset. Challenges such as finger contact precision and occlusions—where instruments like drums can block the performer from certain angles—complicate the full capture of body movements. Finger contact precision, in particular, is a critical yet challenging aspect of digital capture. Musical performance often requires intricate finger movements, such as pressing keys, plucking strings, or covering holes on wind instruments, where even slight variations in pressure or positioning can significantly affect the sound produced. Capturing these micro-level interactions with high fidelity requires precise, high-resolution sensors that can track subtle finger movements in real time. However, the use of specialized equipment, such as motion capture suits or gloves, can hinder the natural flow of performance, limiting the authenticity of the captured data. Nevertheless, the multi-modal nature of musical performance is vital for understanding how musicians interact with their instruments, express emotions, and shape the overall experience of the performance. Being able to capture synchronized audio signals like data from MIDI interfaces can help disambiguate noisy data captured by other sensors and devices.

To address the current gap in data availability and accelerate the research in related topics, we introduce the Multi-Modal Instrument Performances (MMIP) database, a novel dataset that captures comprehensive musical instrument performances through multiple synchronized data streams. The MMIP database offers high-quality 3D motion capture of the musician's body, fingers, facial expressions, and instrument, synchronized with aligned audio, multi-angle video, and MIDI data. The dataset contains a total of 3.5 hours of recordings, featuring performances on three instruments: guitar, digital piano, and drums. The quality of the MMIP database has been evaluated across several metrics, including fingertip-to-key distance accuracy, smoothness of the motion data, temporal fidelity, and audio-to-motion synchronization. These evaluations show the dataset's reliability and robustness, making it suitable for advanced analyses, motion retargeting, and integration into virtual environments. To efficiently manage, organize, and disseminate this complex musical performance data, we have designed and developed an online repository that provides open access to the data along with detailed annotations. The repository includes raw data in formats commonly used across the animation, music, and research communities, ensuring compatibility and ease of use. In addition to the raw dataset, we offer a MIDI-to-instrument animation project, enabling users to visualize instrument performances, and a script for playing back FBX files with audio in a web environment. This functionality allows users to easily preview the data before initiating downloads, streamlining the data exploration process.

The MMIP database offers a wealth of multimodal data with numerous potential applications, significantly advancing research in fields that rely on rich datasets for training deep learning and generative models. This resource is expected to drive developments in character animation, computer vision, and the performing arts, particularly in music. By providing a diverse, multimodal dataset, the MMIP database supports a wide range of research applications, such as audio-to-animation synthesis (e.g., MOSA [HMC*24]), audio-to-MIDI transcription (e.g., Basic Pitch [BBR*22]), and musical instrument performance analysis for injury prevention [ASGA17]. Furthermore, it is a valuable asset for gesture recognition research, where full-body motion capture data enhances the understanding and prediction of musicians' movements. Additionally, the database can aid in the development of intelligent tutoring systems that provide real-time feedback on technique and posture (e.g., piano tutoring [BKD12]), and contribute to the creation of new interactive musical experiences that integrate audio, visual, and motion data. In the context of virtual reality (VR) and augmented reality (AR), the MMIP database could enable the realistic simulation and interaction with virtual instruments, enriching user experiences in immersive environments, thus reshaping the landscape of performing arts. These advancements foster greater inclusivity, creativity, and the possibility of live performances across multiple locations, transforming how performances are conceived and experienced. The potential of digital musical instrument performances is demonstrated by the annual virtual performances of well-known artists, which have gained widespread popularity among audiences, generating significant revenue and audience engagement. Examples include Metallica [For24], 21 Pilots [Mov23], Foo Fighters [Met22], and John Legend [Leg20].

2. Related Work

Recent advances in deep learning and generative models, including those in graphics and animation, rely heavily on the availability of large and diverse datasets. These datasets are crucial for training models that generate realistic motion and interactions, such as BEHAVE [BXP*22], TRUMANS [JZL*24], and ARCTIC [FTT*23]. General-purpose human motion datasets like CMU Motion Capture [CMU03], AMASS [MGT*19], and HUMAN3.6M [IPOS14] have been widely used for pose estimation and motion synthesis tasks. These datasets primarily focus on body poses, without incorporating additional modalities such as audio or text annotations. Several models, such as Phase-Functioned Neural Networks [HKS17], GANimator [LAZ*22], and Motion Diffusion Model [TRG*23] have utilized these pose-only datasets to achieve impressive results in motion synthesis and animation. However, challenges arise when integrating additional modalities, such as audio or text, into the datasets. Multi-modal datasets provide richer information, enabling more complex tasks, like generating movements driven by audio or text inputs. Audio-driven datasets, for instance, synchronize motion with accompanying audio, allowing models to generate movements that respond to music or sound. Notable examples of audio-driven datasets include, among other, the FineDance [LZZ*23], AIST++ [LYRK21], Talking With Hands 16.2M [LDM*19], and DanceDB [ASC19] databases. Such datasets have been instrumental in advancing methods like [AYA*23, ANBH23, CAB*24, XTNK24, JSS*24], which demonstrate the potential of generating expressive motion aligned with audio inputs. Similarly, text-driven datasets, which provide detailed movement annotations, allow models to generate movements based on textual descriptions. These datasets, such as FLAG3D [TLL*23], NTU RGB+D [LSP*20], InterGen [LZL*24a], KIT [PMA16], HumanAct12 [GZW*20], introduce the additional complexity of labeling movements and mapping text to motion. Some examples that uses these type of datasets are [ZHL*24, ZCP*22, ZZC*23]. These methods highlight the critical need for synchronized multi-modal datasets to improve the realism and diversity of generated performances.

Table 1: *Multi-modal Music Performance Datasets*

Name	Instrument	Duration (hours)	Annotation							3D MoCap					
			A	N	M	m	V	EMG	D	U	L	f	I	F	Type
EEP [MRPM14]	String Quartet	N/A	✓	✓		✓							✓		MAG
QUARTET [PMPCM14]	String Quartet	0.9	✓	✓		✓	✓			✓			✓		OPT
Gesture DB [SCTO17]	Piano, Violin	N/A	✓	✓	✓		✓	✓		✓	✓				OPT
TELMI [VKV*17]	Violin	2.4	✓				✓	✓	✓	✓	✓		✓		OPT
MOSA [HMC*24]	Piano, Violin	30.7	✓	✓		✓				✓	✓		✓		OPT
M-M Guitar [PCAW16]	Guitar	0.2	✓	✓		✓	✓			✓		✓	✓		OPT
SPD [JQS*24]	Cello, Violin	3.0	✓			✓	✓			✓	✓	✓	✓		MAR
MMIP	**Guitar, Piano, Drums**	3.5	✓	✓	✓	✓	✓			✓	✓	✓	✓	✓	**IMU, OPT**

A: Audio, **N**: Note, **M**: MIDI, **m**: metadata, **V**: Video, **U**: Upper body, **L**: Lower Body, **f**: fingers, **I**: Instrument, **F**: Facial Expressions, **EMG**: Electromyogram, **D**: Depth, **OPT**: Optical, **IMU**: Inertial Measurement Units, **MAG**: Magnetic, **MAR**: Markerless, ☐ : Restricted, ☐ : Unavailable/Non-working Link, ☐ : This dataset

Despite the progress made in recent years that has enable large-scale data availability, managing multi-modal data continues to present unique challenges. Musical instrument performance, in particular, requires capturing a complex fusion of auditory, visual, and physical features, such as body movements, finger interactions, and facial expressions along with a synchronised audio and MIDI data. Due to this high complexity, most existing datasets either do not capture or fail to effectively integrate these multi-signals, creating a gap in the resources necessary to advance research in musical performance analysis and animation. Currently, most available datasets primarily focus on audio and musical instruments for applications in Music Information Retrieval (MIR) tasks [MSH20, LLD*19, XBP*18, BVGLH17, EBD10, GR06] - for more details, refer to the survey by Kyriakou et al. [KAdlCCP*24]. Additionally, noteworthy datasets such as Motion-X [LZL*24b] and UBody [LZW*23], while including some musical footage, are not explicitly focused on this domain. They lack MIDI or note-level annotations, high-quality audio, and the accuracy of finger-to-note alignment may be imprecise—particularly when derived from a single video source—thus limiting their suitability for detailed musical performance capture. In this work, however, we focus on datasets related to musical instrument performances that include motion capture (MoCap) data, which is crucial for capturing the essence of a musical performance. Currently, only a limited number of such datasets exists, and Table 1 summarizes their data content, instruments, duration, and availability.

Musical Datasets with Limited MoCap Data: The Ensemble Expressive Performance (EEP) Dataset [MRPM14] is one of the early efforts focusing on instrument movements, particularly bowing actions, in string quartet performances using an electromagnetic MoCap system. Although it includes audio tracks and note annotations, it lacks MoCap data for body movements, finger motions, and facial expressions, limiting its focus to instrument motion. This limitation, ignores the performer's gestures that enhance musical expression. Similarly, the QUARTET Dataset [PMPCM14] records upper body and instrument movements using an optical MoCap system, with audio, video, note annotation and bowing descriptors. While valuable for studying the interaction between upper body movements and instrument handling, it does not capture finger movements or facial expressions and lacks lower body motion data, limiting full-body comprehensive analysis.

Musical Datasets Including Upper and Lower Body MoCap: Progressing towards more detailed capture, the Gesture Dataset [SCTO17] integrates upper and lower body motion data for pianists and violinists, offering audio, video, MIDI, and MoCap data for piano performances, along with audio and electromyogram (EMG) readings for violin performances to analyze muscle activation. However, it lacks finger motion capture and facial expressions, which are crucial for fully visualizing and analyzing digital performances. In addition, the dataset's GitLab hosting complicates data exploration due to the absence of features like filtering, searching, and previewing capabilities. Similarly, the TELMI Dataset [VKV*17] includes EMG data, Kinect depth maps, and MoCap for body, violin, and bow, alongside multi-angle video, but it also lacks finger and facial motion data. The MOSA Dataset [HMC*24], the largest of the three, captures full body movements in piano and violin performances, yet, like the others, does not include finger or facial data. Furthermore, it is not open access, which limits its use within the broader research community.

Musical Datasets Capturing Finger Movements: Addressing the critical need for finger recording, the M-M Guitar Dataset [PCAW16] focuses on the motion capture of guitarists' fingers, upper body, and instrument, providing audio and video recordings with note annotations. While offering valuable insights into guitar performances, it is not publicly available, restricting its impact on the research community. Additionally, with only ten musical pieces, the dataset offers a limited scope for research. The String Performance Dataset (SPD) [JQS*24] utilizes a markerless, vision-based MoCap system to capture both upper and lower body movements in cello and violin performances, while also including finger motion data. SPD provides synchronized audio, video, and metadata, enabling comprehensive analysis of string instrument performance, though it lacks facial expression data, limiting studies on the emotional aspects of performance. Additionally, its restricted accessibility, further limits its use in broader research efforts.

In summary, the evolution of musical performance datasets reflects a shift towards more comprehensive data coverage. Early

datasets, such as EEP and QUARTET, focused on instrument movements but do not include body movement data. Later datasets like Gesture and TELMI incorporated upper body MoCap and physiological data, but still missed finer details like facial expressions and finger movements. MOSA, on the other hand, included upper and lower body movements but continued to lack finger and facial data. Datasets like M-M Guitar and SPD advanced by capturing finger movements, essential for instruments like guitar, cello, and violin, yet still excluded facial data, which is closely associated to performer expression. Additionally, as datasets become more multimodal, they tend to have smaller sample sizes and more restricted access. Another key limitation of the existing datasets is that many of them are hosted on platforms not optimized for large data handling, making filtering, searching, and previewing difficult. Most datasets focus on string quartet instruments, with a lack of diversity in instruments. Finally, note annotations are often not provided in MIDI format, and motion capture data is available only in .csv files, which is not the common format used in the character animation industry and not ideal for ready-to-play animations (such as e.g., the FBX or BVH formats).

The MMIP Dataset overcomes these limitations by providing comprehensive MoCap data that includes full-body movements, fine motor skills, facial expressions, along other expressive features. It supports performances on instruments like guitar, digital piano, and drums, and is hosted on a platform optimized for large datasets with advanced filtering, searching, and previewing capabilities. The data is provided in standardized, user-friendly formats like FBX and includes MIDI files for note annotations, allowing immediate use without extensive preprocessing. By offering synchronized audio, MIDI data, metadata, and video recordings, MMIP enables in-depth analysis of the intricate relationship between physical gestures and musical expression. This dataset marks a significant progression in MoCap completeness, setting a new standard for future research in musical performance analysis. Building on such comprehensive and accessible datasets is expected to enable more nuanced analyses, ultimately deepening our understanding of the artistry involved in musical performances.

3. MMIP Dataset

As previously mentioned, the MMIP dataset is a comprehensive multi-modal musical instrument dataset that includes high-quality 3D motion capture of full-body movements, finger positions, and facial expressions of the musician, along with synchronized audio recordings, video footage, and MIDI data for three instruments: guitar, digital piano, and drums. Currently, the dataset has a total duration of 3.5 hours. It is important to emphasize that this is an ongoing effort, and we plan to continually update and enrich the dataset on a regular basis. One crucial aspect of the dataset is the inclusion of MIDI, which plays an essential role in digitizing musical instrument performances. MIDI serves as the ground truth, accurately reflecting what the musicians played by capturing precise details such as note on/off events, dynamics, and time signatures. This enables the exact replication of performances, providing a level of accuracy that is invaluable for evaluation, post-processing, cleaning, and revising the data. Moreover, the use of MIDI significantly reduces the time and effort required compared to manual note annotation. Manual annotation is not only time-consuming but also prone to errors, making MIDI a critical tool in both music analysis and production, ensuring efficiency and precision.

Figure 1: *The studio where our captures took place.*

3.1. Capturing Devices

To digitize such a multi-modal musical performance, we require an array of devices capable of capturing both body movements and audio recordings in a synchronized manner. This section provides a detailed overview of the equipment used in our experiments. The recordings were conducted in our dedicated studio (as shown in Figure 1), which has a 100-square-meter capture area, a 30-square-meter LED wall capable of displaying highly accurate colors at an extremely high resolution with low latency (less than 25ms), Dolby surround audio speakers, and controlled lighting conditions.

For full-body motion capture of the musicians, we employed the Rokoko Full Performance Capture system [Rok24b]. This system includes a suit for tracking body movements, gloves for monitoring finger movements, and an iPhone 12 Pro Max for capturing facial expressions. Rokoko operates on an inertial motion capture framework, which, despite certain limitations—such as potential data drifting and inaccuracies in 3D data representation—proves suitable for specific use cases. In the context of capturing musical instrument performances, particularly with larger instruments like drums that often lead to occlusions, inertial motion capture is preferred over optical systems. Optical systems, though highly accurate in many scenarios, frequently encounter challenges when capturing performances involving occlusions, as these can significantly impact the accuracy of the data. In contrast, inertial systems like Rokoko, which rely on sensors embedded in the suit and gloves, are better equipped to handle such environments, making them a more reliable choice for capturing complex performances with minimal data interference. To enhance the precision of motion data and eliminate drifting during live performances, we additionally employed the Rokoko Coil Pro system [Rok24a], which utilizes electromagnetic field (EMF) technology. This technology allows for accurate tracking of the hands' absolute positions in space, effectively preventing issues like hand intersections. To facilitate the alignment of cameras and the reconstruction of 3D data,

Figure 2: *The musician, dressed in a Rokoko MoCap suit, performs on a digital piano. "A" shows the coil pro (the right box in the tripod), used with the gloves to enable accurate finger capture. "B" shows the headrig used for capturing facial data with an iPhone. "C" displays the ChArUco board placed on the floor to facilitate camera parameter calibration.*

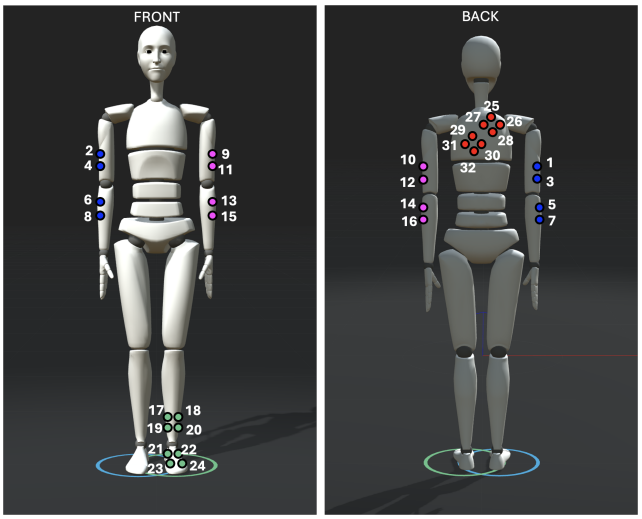

Figure 3: *PhaseSpace body markers' placement. On the left is the front view, and on the right is the back view. Markers were placed as follows: 4 markers on each upper arm (1–4, 9–12) and lower arm (5–8, 13–16), totaling 16; 4 markers on the left lower leg (17–20); 4 markers on the left foot (21–24); and 8 markers on the back (25–32).*

a ChArUco board was placed on the floor as a calibration reference. The board, combining checkerboard patterns with ArUco markers, provides precise reference points for both intrinsic and extrinsic camera calibration. Figure 2 shows a snapshot with the capturing configuration. In the case of guitar recordings, we further enhanced the setup by using a 24-cameras Phasespace Impulse X2E motion capture system with active markers, alongside the Rokoko system, to precisely track the guitar's movement in space. This dual-system approach ensured a higher level of accuracy and synchronization between the musician's movements and the instrument's spatial positioning.

Audio and MIDI data were captured using REAPER [Coc24], a digital audio workstation (DAW). For electronic instruments such as keyboards and drums that have audio and MIDI outputs, we directly connected the instruments to the computer. For electric and acoustic instruments like the guitar, which lack direct audio and MIDI outputs, we utilized the Focusrite Scarlett 2i2 [Foc24] audio interface to capture high-quality audio. Additionally, we employed audio-to-MIDI plugins to obtain MIDI data from these instruments. One such plugin is NeuralNote [RVM24], which offers advanced audio-to-MIDI conversion capabilities within a DAW. NeuralNote is compatible with any tonal instrument, including voice, supports polyphonic transcription and pitch bend detection, and is characterized by its lightweight design and rapid transcription speed. Internally, NeuralNote utilizes the model from Spotify's Basic Pitch [BBR*22].

Finally, our setup included five strategically positioned RGB vision cameras to record video footage from multiple angles, ensuring a comprehensive reference of the original performance. This multi-camera configuration not only provided a detailed visual record but also enabled vision-based applications, evaluation, and analysis, while enhancing synchronization of multi-modal data. Specifically, we employed two Apple 12.9-inch iPad Pro 3rd Generation devices, one Apple 11-inch iPad Pro 7th Generation, an Insta360 X4 camera, and a Samsung Galaxy S21 Ultra. The iPads were chosen for their high-resolution video capabilities and consistent frame rate performance, while the Insta360 X4 was selected for its high-fidelity captures and ability to provide immersive views of the performance environment.

3.2. Apparatus Settings

The apparatus setup for full-body motion capture employed the Rokoko Full Performance Capture system, configured to record at 60 frames per second (fps). This frame rate is sufficient to capture even the fastest movements of the performer, including the fine details of finger movements and face (note that all capturing devices used in our recordings were configured to operate at 60 fps). For stationary instruments like the digital piano and drums, no motion capture devices were required. Note that the drumsticks are not stationary, and attaching a tracking device to them presents a challenge, primarily due to the rapid motion and impact forces generated when they strike the drum. These factors make it difficult to securely affix a tracking device without compromising performance, risking damage, or producing noisy data. Therefore, the positions of the drumsticks can be inferred using the contact point on the drum and the hand movements of the performer. For moving instruments, such as the guitar, since the Rokoko Full Performance Capture system does not include sensors for props, we utilized the PhaseSpace Impulse X2E optical motion capture system,

Figure 4: *PhaseSpace guitar markers' placement. On the left is the front view of the guitar, and on the right is the back view. Four markers (1–4) were placed on the front of the head, four on the back of the head (5–8), and eight markers on the body (9–15).*

which also operates at 60 fps. A total of 24 markers were placed on the performer's body (as shown in Figure 3) to ensure proper association and alignment between the data from the two motion capture systems, Rokoko and PhaseSpace. Additionally, 16 markers were affixed to the guitar (see Figure 4) to track the instrument's movements in space accurately.

Room audio for the performances was captured using a high-resolution stereo microphone from Focusrite, which was recorded through the Digital Audio Workstation (DAW) REAPER at a sampling rate of 44,100 Hz. The instruments' sound was captured directly through the Scarlet 2i2 audio interface. For the drums and digital piano, we simultaneously recorded the MIDI data directly from the instrument onto a separate channel in REAPER, ensuring perfect synchronization between the audio and MIDI data.

The video recordings were captured using four cameras on tripods positioned diagonally around the performer at a height of 1.4 meters and a distance of 3 meters from the performer, ensuring multiple angles of the performance were recorded. Each camera is positioned to face the performer, with subsequent cameras angled at an additional 90°, starting from an initial angle of 45°. Detailed specifications for each camera, including their exact positions and technical settings, are provided as metadata on a performance-by-performance basis on the website. Additionally, for performances involving drums and digital piano, an extra overhead camera was placed directly above the performer at a height of 2 meters. All cameras recorded in 4K resolution at a frame rate of 60 fps, ensuring high-definition video suitable for both performance review and detailed visual analysis.

3.3. Privacy considerations

A careful and detailed examination of the privacy and ethical implications associated with the application of these methods is essential. The digitization of an artist and their performance necessitates the collection of sensitive personal data, such as movement patterns, playing techniques, and biofeedback information [KAdlCCP*24]. Given the sensitive nature of this data, it is imperative to ensure that ethical guidelines are strictly adhered to. To mitigate privacy concerns, informed consent was obtained from all participating musicians (consent forms were given in the local language, and can be provided upon request). Each musician provided written consent after being fully briefed on the study's objectives, the data being collected, and how it would be used. In addition to consent, participants were compensated with 150 euros for their time and contribution to the research, further ensuring transparency and ethical compliance.

3.4. Performers

Our recording sessions featured four highly skilled professional musicians, each with extensive experience and academic credentials in their respective fields, which contributed to the robust foundation of our dataset. These musicians bring a rich diversity of expertise across jazz, classical music, composition, and percussion, fostering a comprehensive exploration of various musical instrument performances. Their advanced qualifications, combined with years of professional experience, were crucial in ensuring the high quality of the recorded dataset. The quality of the produced music relied heavily on their deep understanding of music theory, technical proficiency, and improvisational abilities. A brief biography of each performer, highlighting their backgrounds and qualifications, can be found in Appendix A.

3.5. Musical Instruments

The selection of musical instruments for this study was made with careful consideration of their audio quality, expressive capabilities, and technical characteristics. We selected the following musical instruments to initiate our database due to their global popularity among musicians, the desire to include a diverse range of stringed and percussion instruments in our dataset, and their ability to support MIDI information, among other factors.

Digital piano: It is a digital keyboard instrument designed to emanate the sound, feel, and experience of an acoustic piano. We used the ROLAND FP-10, a full 88-key weighted keyboard, providing musicians with a realistic playing experience. Additionally, it features a built-in USB MIDI interface, allowing for seamless connection to DAW software.

Drums: For the drums, we selected the Carlsbro CSD500 8-Piece Electronic Mesh Head Kit, which includes five mesh drum pads (an 8" bass pad with kick pedal, a 10" dual-zone snare, and three 8" dual-zone toms). The kit also features three cymbals: one 10" single-zone cymbal and two 12" dual-zone cymbals. It is equipped with a USB MIDI interface and MIDI IN/OUT, allowing for versatile connectivity and high-quality audio capture.

Guitar: For the guitar, we used the PRS SE Paul Allender electric guitar that features a 24-fret wide-thin maple neck with a rosewood fretboard and jumbo frets, EMG pickups, coil-tapped volume and tone controls, and a scale length of 25".

Figure 5: *Performance Recording: The top section shows the drum recording, the middle features the digital piano, and the bottom displays the guitar recording. On the left side, there is a top-down view of the drums and digital piano, and the optical MoCap data for the guitar. Four snapshots from the video recording, taken from different camera angles, are also presented for each instrument. On the right, a 3D motion reconstruction is displayed, using the default actor model from Rokoko Studio software to accurately represent the performer's movements.*

4. Data Capture Protocol

In this section, we detail the data collection procedures employed during our recording sessions. In addition, we explain how different modalities and signals were synchronized to ensure accurate alignment, the process for collecting associated metadata for each musical performance, and the variety of data formats captured and curated. Additionally, we describe the editing steps taken to refine the data before it was stored and how these various data formats are shared within our database.

4.1. Data collection procedure

Before collecting the performance data, we thoroughly discussed and explained the project's objectives with the musicians, ensuring they were fully informed of the ethical considerations involved in the research. Prior to their scheduled session at the recording studio, each musician received comprehensive documentation detailing the study's procedures, including information on data collection, the use of motion capture technology, how the data would be utilized, and their rights as participants.

Upon arrival at the recording studio, the musicians were asked to review and sign an informed consent form, which confirmed their understanding of the study's purpose, their voluntary participation, and their agreement to adhere to the ethical guidelines established by the research team. Once the consent process was completed, we conducted a series of body measurements for each musician, encompassing 11 specific dimensions as guided by the Rokoko specifications [Rok23]. These measurements were essential to ensure the accuracy of the motion capture data.

Following this, the musicians wore the Rokoko motion capture suit, which included specialized gloves and a headrig for comprehensive full-body tracking. Before data collection began, each musician was given time to warm up and familiarize themselves with the motion capture suit. This step was crucial to ensure they felt comfortable, confident, and fully at ease with the suit's components, minimizing any potential restrictions to their movement during the performance. Figure 5 illustrates the data collection procedure.

Before each recording session began, detailed instructions were provided to the musicians regarding the calibration of the motion capture equipment. This calibration process was crucial to ensure accurate tracking and was performed regularly throughout the session to maintain the highest data fidelity. For the pianists and drummer, we chose not to use the optical MoCap system to track the positions of their instruments, as we did for the guitar, since these instruments are static and do not require continuous positional tracking. Instead, we employed a predefined calibration procedure before the start of each recording track to accurately register the instrument positions in relation to the musicians. For the pianists, this calibration involved simultaneously pressing both the first and

last keys on the piano, allowing us to establish the instrument's exact spatial orientation. This simple yet effective method provided a fixed reference for the motion capture system to align with during the performance. Similarly, for the drummer, the calibration process involved a sequence of strikes to map the drum set in space. The drummer first played the bass drum, followed by simultaneous hits on the snare drum and floor tom, and then hit the hi-tom and mid-tom together. This sequence provided clear spatial markers for each drum, ensuring that the motion capture system could accurately track the drummer's interactions with the entire set.

During the recording sessions, the musicians performed either to a metronome or backing tracks to ensure rhythmic precision. It is important to note that the metronome, a widely used tool in the music industry, produces a regular click sound at set intervals, helping musicians maintain a consistent tempo by providing a reference in beats per minute (BPM). This prevents tempo drifting, ensuring that the performance remains tightly aligned with the desired rhythm. In our setup, we utilized the digital metronome within the REAPER DAW for this purpose. This allowed for easy synchronization between the audio and motion capture elements of the recording process. Note that all performers received audio feedback through an amplifier as they performed, which is a common practice when playing a musical instrument.

To prevent any potential data drift in the motion capture system, particularly as per Rokoko's guidelines, we limited the duration of each recording to a maximum of five minutes. This precaution was essential to ensure that the motion capture data remained accurate and reliable throughout the performance, as extended recording sessions may lead to minor drifts in data integrity over time.

4.2. Data Synchronization

Synchronization is a critical but challenging aspect in multi-modal setups, especially when dealing with various data streams. In our case, we needed to synchronize motion capture data, in some cases from two different systems, room audio, instrument audio, MIDI data, and video recordings. Ensuring precise alignment between these different modalities was essential to maintain the integrity of the collected performance data.

Full-body MoCap Data and Audio Synchronization: To synchronize the motion capture data with the audio, room audio was recorded simultaneously using a Digital Audio Workstation (DAW) alongside the motion capture system. Both the audio and MIDI were captured within the same DAW project on separate channels, ensuring precise synchronization. For full-body alignment with the audio, the Rokoko motion capture software and REAPER DAW were started in sync using a Python script, ensuring that all motion capture data, audio, and MIDI signals were consistently aligned (see Figure 6 - part a).

Audio and Video Synchronization: For synchronizing video with audio, we used DaVinci Resolve's automatic alignment tool (Figure 6 - part c), which matches the audio waveforms captured by the video cameras (Figure 6 - part b) with the room audio from the REAPER DAW (Figure 6 - part a). This method enabled precise alignment, with no noticeable discrepancies during playback.

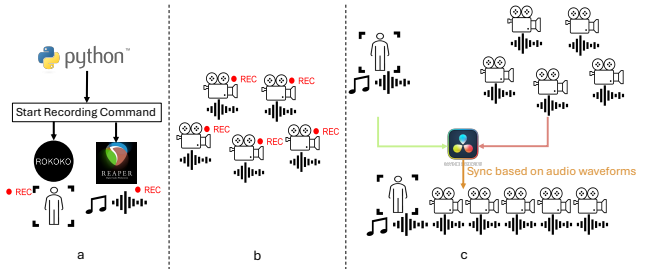

Figure 6: *Synchronization Process: a) A Python script was developed to start audio, MIDI and motion capture recordings simultaneously. b) Multi-angled asynchronous videos were recorded. c) Room audio serves as the reference for synchronizing all data sources using audio waveforms.*

Visual markers, such as metronome spikes in the audio, further supported the synchronization process. Despite the accuracy of the automatic synchronization, we performed manual verification to ensure perfect alignment between the motion capture data and video. This involved reviewing the video footage alongside the motion capture data, frame by frame, in Blender. If any misalignment was detected, manual offsets were applied to correct the timing, and the corrected data was then re-exported from Rokoko Studio.

Instrument Position and MoCap Synchronization: In contrast, synchronizing the optical motion capture data of the guitar with the full-body motion capture data required manual effort in specialized animation software like MotionBuilder. This process involved frame-by-frame alignment and typically took around five minutes per recording. To ensure precise synchronization, musicians were asked to perform distinct, quick movements at the beginning of each take. For instance, the guitarist initiated a foot tap, while the pianists combined a foot tap with pressing the first and last keys of the piano. The drummer's synchronization was aided by a sequence that started with a bass drum kick, followed by hits on the snare, floor tom, hi-tom, and mid-tom. Note that, the drummer's rapid movements made synchronization easier, although there were challenges. In some recordings, the intensity of the drummer's hits caused disruptions in the motion capture sensors, leading to data loss. Only intact recordings were used, and in a few cases, the drummer's global position had to be manually edited to correct drift over time.

4.3. Data annotation and Metadata Information

Metadata is data that provides information about other data; in other words, it serves as the documentation that describes and situates the data within its proper context. Metadata is crucial because it organizes, contextualizes, and enhances the searchability of data, making it easier to find, manage, and comprehend. It ensures data accuracy, tracks usage, facilitates deeper content analysis, and supports efficient data integration and automation. In short, metadata adds structure and meaning to data, improving its usability and integrity.

In this work, we have comprehensively recorded all relevant data

associated with data collection and processing, along with administrative metadata; this includes recording date, file type, performer details, location, systems used, authentication methods, access controls, and administrative rights. Performer-specific metadata, including name, age, years of experience, education, and body dimensions, has also been documented. This holistic approach provides rich context for anyone analyzing the performance, enabling a deeper understanding of both the musical content and the conditions under which the data was captured.

Furthermore, we included specialized music metadata, focusing on key musical features such as the genre of the piece, time signature, and instruments used. Additionally, we recorded MIDI data, which contains all the critical information needed to understand each musical performance in detail. MIDI captures changes in dynamics, specific notes, and chords played, as well as instrument-specific events like pedal presses for pianists or drum hits for drummers, tagging each action with precise timing and intensity. It is important to note that while MIDI data can be automatically exported for digital pianos and drums, this is not the case for guitars. However, we included guitar MIDI data by using audio-to-MIDI conversion software, specifically Ableton. Thanks to the high quality and clarity of our recordings, this transformation produced high-quality MIDI data, ensuring consistency across the dataset.

4.4. Data Formats

In this subsection, we present the various data types captured during the performances, including motion, audio, and video recordings. Each data format was carefully chosen to ensure high fidelity, detailed representation, portability, and transparency. Our selection follows industry standards and ensures compatibility with widely used software, enabling comprehensive analysis.

Motion Capture Data: The motion capture data is provided in FBX format at 60 frames per second (fps). The FBX format is a file format developed by Autodesk, and supports complex data like 3D models, animations, textures, lighting, and motion capture, making it popular in game development, animation, and visual effects. In our dataset, we used the default human actor model (see Figure 5, right, for an illustration of the avatar) exported from Rokoko Studio software, which supports full-body and finger articulations, as well as facial expressions. Notably, the actor model used in the animation accurately reflects the real-life dimensions of each performer, ensuring high fidelity in movement representation. This accuracy is achieved by creating the model using 11 body measurements from the actual performer, so each exported FBX file maintains the correct proportions of the performer.

Guitar Tracking: For guitar performances, motion tracking was captured using PhaseSpace and is available in the C3D format, offering detailed spatial information about the guitar player's movements. C3D is a standard file type to store 3D positional data of markers and associated analog data, and includes metadata like sample rates and calibration information.

Audio Recordings: Audio recordings of the performances are available as stereo tracks in WAV (Waveform Audio File) format, recorded at a sampling rate of 44.1 kHz. WAV is a standard audio file format used to store raw, uncompressed audio data. This format ensures high-quality audio suitable for detailed analysis.

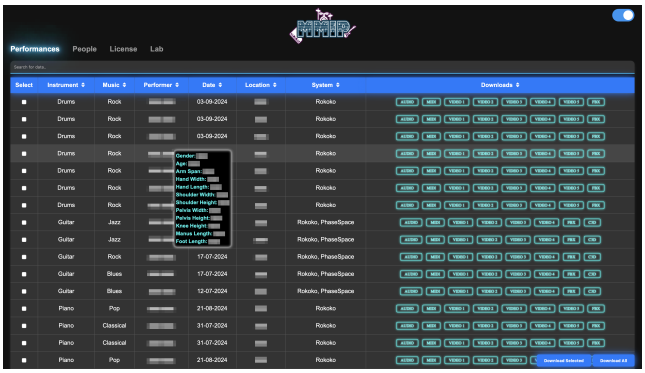

Figure 7: *The MMIP Website: the website can be accessed via the following link: https://mmip.cs.ucy.ac.cy.*

MIDI: MIDI data is provided in MID format, capturing detailed information about notes (pitch), chords, dynamics, velocity (volume), timing, control changes (like modulation or sustain), tempo, and instrument-specific actions during the performances. It does not contain actual audio but instead provides instructions for playing the music, making the files compact and versatile for music production and composition.

Video Recordings Videos were shot from multiple angles in 4K resolution at 60 fps and are available in MP4 format using the H.264 codec. MP4 is a widely-used digital multimedia format for storing video, audio, and data such as subtitles. Its popularity stems from its high compression efficiency, which preserves quality while keeping file sizes manageable. These high-definition recordings provide a comprehensive visual perspective of each performance, complementing the motion capture and audio data.

5. The Online Repository and Interface

To ensure easy access and navigation in a user-friendly environment, we have designed and developed an open-access online repository for the collected performance data. As shown in Figure 7, the interface is intuitive and simple to use, featuring a variety of built-in visualization tools that enable real-time exploration of the dataset. Users can seamlessly view motion capture animations, listen to synchronized audio recordings, and watch performance videos from multiple angles, all through interactive visualizations hosted directly on the platform. This comprehensive interface enhances the accessibility and usability of the data, allowing researchers and users to fully engage with the content without the need for external software.

5.1. Accessibility and Interaction

Our repository interface has been designed to provide a range of features that enable users to easily navigate, locate, and retrieve the precise data they need, greatly enhancing both accessibility and interactivity. In particular, users can choose to download individual

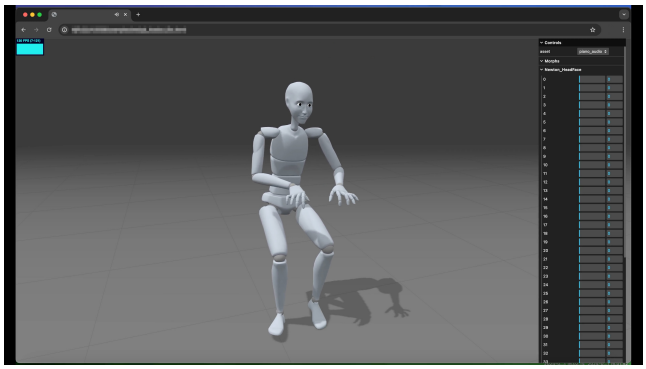

Figure 8: *Visualization of our custom FBX and audio web player.*

files, specific performances, or the entire collection. The platform also includes powerful search functionalities that allow for efficient filtering and sorting based on metadata. Users can organize the data according to various parameters, such as instrument, performer, music genre, recording location, or the capture systems used. Figure 7 illustrates the website interface; for a detailed demonstration of all functionalities, including the search tools, please refer to the accompanying video. Moreover, hovering over specific cells in the interface reveals additional metadata, such as the performer's body dimensions and key musical features of the piece. This functionality provides users with quick access to detailed information, further enriching their exploration of the dataset.

In addition to data access, the platform provides in-built tools for easy visualizations, enhancing accessibility prior to downloading the data. For instance, when users click on a video, it opens in a new window where playback begins automatically, offering a seamless viewing experience. Similarly, selecting an FBX file opens a new window featuring a 3D visualization of the performance, synchronized with audio (see Subsection 5.2 for further details on the custom tools provided). Users can navigate freely within the 3D environment, allowing them to view the performance from any desired angle. A representative screenshot of this interface is shown in Figure 8.

5.2. Tools Provided

To further facilitate the exploration and analysis of the dataset, we provide additional tools alongside the data:

FBX Player Script: A custom script is included in the repository interface that allows users to playback FBX files synchronized with audio directly in a web environment, eliminating the need for external 3D software or downloading the data. This feature provides flexibility in data exploration, offering a comprehensive view of the performers' movements and interactions with their instruments. Users can interactively explore the motion capture data and audio, with controls that allow navigation within the 3D scene and adjustment of the view for more detailed analysis.

MIDI-to-Instrument Animation Project: This tool enables the 3D visualization of musical instruments by generating animations based on MIDI data. By accurately reflecting the notes played by musicians during the performance, it allows animators or researchers to determine precise finger positions on the instruments. This data can be integrated into automated systems, such as Inverse Kinematics (IK) controllers, to align finger movements with the correct notes in the animation, producing highly realistic and detailed representations of the performance. As shown in Figure 9, highlighted notes indicate the active keys during the piano and drums performance and Figure 12 illustrates an example of post-processing improvements in fingertip-to-key accuracy achieved using a standard IK solver.

Figure 9: *Piano and Drums MIDI players: Purple notes indicate the white keys on the piano, and gold notes indicate the black keys. A similar color scheme is used for the drum kit, with gold representing the cymbals.*

5.3. License

All content in the repository is released under the Creative Commons license (CC BY-NC-SA 4.0) [Com24], allowing for free sharing, adaptation, and use of the material for noncommercial purposes only, provided that proper credit is given to the original authors. If you remix, transform, or build upon the material, you must distribute your contributions under the same license as the original. The license is clearly displayed in the repository, ensuring that users acknowledge and accept it before downloading the data.

6. Dataset Evaluation

This section evaluates the quality and robustness of the captured dataset, focusing on two key components: motion accuracy and

motion quality. Motion accuracy examines the precision of finger-to-key alignment in piano performances, while motion quality assesses the overall reliability and smoothness of the dataset, with particular emphasis on synchronization and temporal fidelity. The primary objective of this evaluation is to verify the accuracy of the dataset, ensuring it meets the high-quality standards necessary for seamless integration into virtual environments. By presenting quantitative metrics, we aim to offer a comprehensive understanding of the dataset's strengths and identify potential areas for improvement.

6.1. Motion Accuracy

This section evaluates finger position accuracy in piano performances. Unlike guitars, where a single finger can press multiple notes (e.g., barre chords), and drums, where positional precision is less critical than timing and force, piano playing involves discrete key presses with direct fingertip-to-key interaction. This makes the piano an ideal candidate for precision analysis, as the spatial relationship between fingers and keys directly impacts note accuracy and is straightforward to measure. While the focus is on the piano, similar accuracy levels are anticipated for the other two instruments.

Before evaluating the precision of piano performances, it is essential to establish a robust methodology for mapping the physical interaction between the performer's hands and the instrument. This involves three important steps: (1) defining the spatial alignment between the virtual piano keys and the motion-captured hand movements, (2) determining which notes are active at each frame based on MIDI data, and (3) calculating the distances between fingertips and the corresponding keys for every note played. By combining these processes, we can quantitatively assess the accuracy and efficiency of the performer's hand movements with respect to the intended notes. The following paragraphs outline the methodology used to achieve these objectives. Each step leverages a combination of MIDI data for timing and note identification, made with a script in Blender.

Piano Alignment: To align the virtual piano with the performer's hand, the script uses two notes—"A-1" and "C7"-representing the lowest and highest keys of an 88-key piano. First, the script parses the MIDI data to locate all "note on" events, mapping each note onset to a corresponding frame in Blender's timeline. It then searches for the first frames where A-1 and C7 occur (in every recording, pianists pressed the first and last note simultaneously, as mentioned in subsection 4.1) and marks these as reference frames for alignment. At these frames, the script retrieves the 3D world-space position of the specific active fingertip bones to define the start and end points of the piano (Figure 10).

After obtaining these two fingertip locations, the script calculates the positions of all 88 keys by first assigning local offsets along a single axis (with white keys spaced at 2.3cm and black keys placed at halfway intervals). It then scales these offsets to match the 3D distance between the two captured finger positions, thereby ensuring that the virtual piano is stretched or compressed to align precisely with the performer's physical notes. This scaling is performed by computing the total white-key distance, comparing it to the measured distance between the fingertips, and deriving a scale

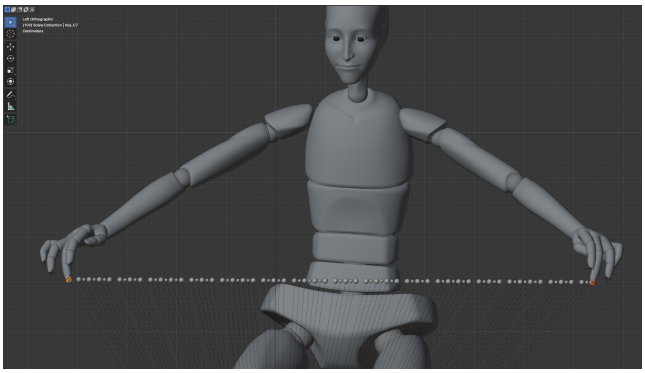

Figure 10: *Visualization of the piano alignment process using A-1 (the lowest key) and C7 (the highest key) as anchor points. The captured fingertip positions at each key press serve as references to scale and orient the entire 88-key layout in 3D space, ensuring that subsequent finger-to-key distance measurements accurately reflect the performer's hand movements.*

factor. The scaling is beneficial for retargeting purposes, particularly when characters with varying size proportions are used, as it allows the virtual piano to be appropriately scaled up or down to accommodate different performers. The entire keyboard is then rotated to follow the vector from the leftmost note (A-1) to the rightmost note (C7). This vector also serves as a reference axis for computing parallel distances in subsequent steps.

Caching Fingertip Positions: Once the piano is aligned in 3D space, the script processes each frame within the relevant time span of the MIDI performance. It iterates from the earliest frame in the note data to the latest, retrieving and storing fingertip positions for both hands. These data are collected in a dictionary mapping every frame to a set of fingertip positions, ensuring that distance calculations can be made efficiently for any note at any moment.

Distance Calculation between Fingers and Keys: For every note onset frame, it identifies which keys become active and calculates the distance for each fingertip to that key's 3D position. Using a projection onto the piano's alignment vector, the script determines if the fingertip is within half the key's width (1.15cm for white keys; 0.5cm for black keys); if so, the distance is treated as zero (whether the note is hit at the center or the edge, it is still considered the same note). Otherwise, it returns the remaining distance beyond that boundary. For instance, when the fingertip is positioned 2 cm away from the center of a white key, which has a width of 2.3 cm, the offset is calculated as $2.0 - \frac{2.3}{2} = 0.85$ cm. Similarly, for a black key with a narrower width of 1.0 cm, the same fingertip distance of 2 cm results in an offset of $2.0 - \frac{1.0}{2} = 1.5$ cm. By comparing all fingers, the script identifies which finger is closest to the key at onset, and records that distance.

Conclusion and Discussion of Results: In total, 40,063 notes from all piano recordings were analyzed using the methodology described above. The distance calculations revealed that the minimum distance between a fingertip and the corresponding key could be as

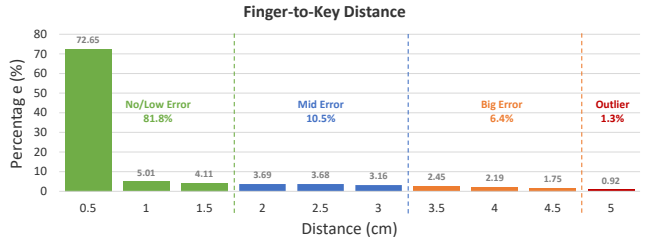

Figure 11: *Histogram showing the distribution of distances (in cm) for piano notes, grouped in 0.5 cm bins. The vertical axis represents the percentage within each distance range, providing insights into the frequency of various distances.*

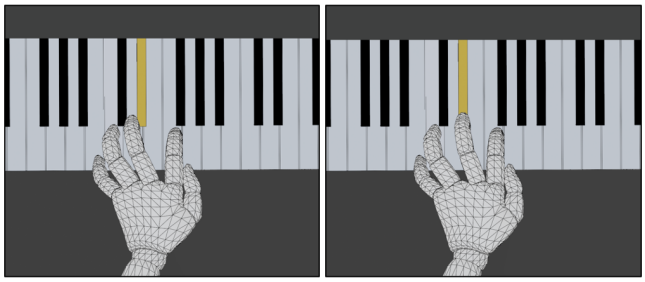

Figure 12: *Post-processing improvements in fingertip-to-key accuracy using MIDI information and Blender's default IK solver. The initial hand pose (left) shows a 1.1cm misalignment between the middle fingertip and the D#2 key. Post-processing (right) corrects this, ensuring the fingertip presses the correct key (highlighted note from MIDI data).*

low as 0.0cm, indicating several instances where fingers were effectively "on" the key at note onset. The average distance of 0.6cm indicates that, on the whole, the performer's fingers remained quite close to the intended keys, reflecting a generally high level of precision. This is consistent with the median distance being 0.0cm, which implies that more than half of the note onsets occurred with the fingers effectively on target. A detailed look at the distance distribution (see Figure 11) shows that 81.8% of onsets fell below 1.5 cm, another 10.5% were between 1.5 cm and 3.0 cm, 6.4% were between 3.0 cm and 4.5 cm, and the remaining 1.3% were greater than 4.5 cm from the ideal key centers. The highest distance recorded was 6.6 cm, which was an outlier as it occurred only once.

The analyzed piano performances comprise approximately 2 hours of recorded material, with 90.5% of the notes exhibiting nearly perfect finger-to-key positioning (81.8% no/low error + 10.5% mid error), corresponding to approximately 1 hour and 48 minutes of precise or near-precise alignment. These results highlight the dataset's reliability in capturing finger-to-key alignment and provide valuable insights into hand positioning across extended performances. While minor deviations may reflect performance nuances, the overall accuracy underscores the quality and utility of the recorded data.

While the dataset demonstrates high quality, it can be further refined through post-processing techniques. Its multi-modality nature enables enhancements such as precise finger positioning on instruments using MIDI data and inverse kinematics, which allow for retargeting to virtual instruments of varying sizes. Similarly, the data can be adapted to virtual characters with diverse body proportions, ensuring realistic and natural motion. These capabilities enhance the dataset's flexibility, making it highly adaptable for a wide range of research and application scenarios. Figure 12 demonstrating the fingertip's offset relative to the piano key both before and after post-processing using a simple IK solver. This figure highlights the significance of MIDI information in enhancing data quality during post-processing by leveraging precise knowledge of the exact key pressed.

6.2. Motion Quality

This evaluation aims to assess the quality and robustness of the dataset through quantitative analyses, focusing on key aspects such as synchronization accuracy and data smoothness. A random sample from the dataset was analyzed to ensure an unbiased assessment. By examining these metrics, we aim to establish the dataset's reliability.

Synchronization: We detected key press events as minima in the Z-axis bone position data using a peak-finding algorithm and matched these with audio onsets detected via Librosa's onset detection algorithm. The time offsets between these events yielded an average synchronization offset of 0.008 seconds, demonstrating excellent synchronization accuracy. This aligns with the natural delay between physical note playing and sound production.

Smoothness: The quality of the captured data required no corrections, as reflected by the following metrics:

1. Temporal Fidelity: Frame intervals confirm a consistent frame rate of 60 fps, ensuring reliable data capture, particularly for rapid finger movements.
2. Signal-to-Noise Ratio (SNR): Our data have a high SNR score (up to 97dB), outperforming other high-quality datasets such as MOSA (88dB) and Gestures (76dB). SNR is calculated by comparing the signal power of smoothed motion data to the noise power derived from deviations between the raw and smoothed data.

Certain quality metrics, such as foot sliding or variety, do not apply to our dataset since musicians were either seated (piano and drums) or standing in fixed positions (guitar), minimizing foot movement compared to dynamic activities like dance or locomotion.

7. Discussion and Conclusions

In this work we introduce an innovative dataset for musical instrument performances. The dataset is distinguished by its comprehensive integration of motion capture data, providing a detailed and synchronized view of a musician's body movements, finger motions, and facial expressions, alongside MIDI, audio, and video data. To the best of our knowledge, this is the first publicly available dataset to combine such a rich array of data points, offering researchers an unprecedented opportunity to study musical performance in a holistic manner.

7.1. A Comprehensive Dataset and Applications

The primary contribution of this research is the creation of a publicly accessible dataset that integrates diverse modalities. This dataset fills a gap in the research landscape, offering a comprehensive view of the various physical and auditory elements involved in a musical performance.

By making this dataset publicly accessible, new doors are opened for important research across various fields, such as in musicology, where users can analyze performance practices and stylistic nuances in ways that were not previously possible; in motion analysis, where researchers can study how musicians' movements are tied to their musical output; in virtual reality and entertainment industry, where uses can use data for visualization purposes; and in deep learning, where our dataset could provide a solid foundation for developing and training advanced deep learning models capable of analyzing, replicating, and synthesizing music performances of multiple modalities. This will allow researchers to explore more sophisticated models that can interpret and reproduce the nuanced expressions and gestures captured in the data. This dataset not only serves as a foundational resource for current research but also paves the way for future investigations into the intersection of music, motion, and machine learning.

One of the main challenges we encountered was maintaining the quality of the animation, particularly the precision of contact points during finer finger movements. This precision is crucial, especially when the correct placement of the fingers significantly impacts both the quality and tone of the produced sound. As shown in our evaluation section, while our motion capture system generally achieves sufficient fidelity to reconstruct subtle body and finger movements, there have been instances where the system fell short in capturing the desired accuracy, resulting in some outliers. However, as illustrated in Figure 12, our dataset includes MIDI information that encodes which notes are pressed. This allows finger precision to be restored during post-processing using IK, thereby improving the accuracy of the contact points, ultimately increasing both the realism and the overall quality of the final output.

7.2. Future Work

This is an ongoing initiative, and we are committed to continuously uploading new data while maintaining the integrity of the current database. Additionally, we plan to establish a standardized uploading protocol that will guide other users in contributing data to our repository. This protocol will include best practices for data submission, offering recommendations on model usage, data representation, and acceptable data formats, among other key factors.

Our long-term objective is to expand the repository's size and scope to ensure greater diversity within the dataset. This includes addressing a wide range of musical attributes, such as rhythm, tempo, and the subtle nuances present in various performances. We also aim to account for diversity in the instruments used and the performers themselves. By taking these factors into consideration, we strive to create a dataset that is not only comprehensive but also truly representative of the rich spectrum of musical expressions across different styles, traditions, and cultures.

7.3. Ethical Considerations and Privacy Concerns

Ethical and privacy issues are central to the development and use of this dataset. The process of digitizing an artist's performance captures sensitive data, including their unique motion patterns and playing style, raising significant concerns about how this information may be used. The potential for unauthorized use of a performer's data to generate new, synthesized performances poses critical questions about the ownership of the resulting content. These issues mirror ongoing discussions about generative models in domains like image and audio synthesis [Bar23], as well as large language models [WMR*21,Har23], which similarly deal with the use of personal data in machine-generated outputs. To address these concerns, each participant in the dataset collection process provided informed consent, fully aware of the scope and purpose of the research. Additionally, the study received approval from the relevant institutional review boards, ensuring adherence to the highest ethical standards. These steps help mitigate privacy risks, but they also highlight the broader ethical challenges that come with applying deep learning and extended reality (XR) technologies to sensitive data. The field of ethics in Artificial Intelligence continues to grow in importance; for additional insights into these ethical issues, the work by Slater et al. [SGLH*20] provides a comprehensive examination of the challenges posed by such emerging technologies.

Acknowledgements

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under Grant Agreement No 739578 and the Government of the Republic of Cyprus through the Deputy Ministry of Research, Innovation and Digital Policy. This project is funded by the European Commission's Horizon Europe program under grant agreements 101178362 and 101061303. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

References

[ANBH23] ALEXANDERSON S., NAGY R., BESKOW J., HENTER G. E.: Listen, denoise, action! audio-driven motion synthesis with diffusion models. *ACM Trans. Graph. 42*, 4 (July 2023). doi:10.1145/3592458. 2

[ASC19] ARISTIDOU A., SHAMIR A., CHRYSANTHOU Y.: Digital dance ethnography: Organizing large dance collections. *J. Comput. Cult. Herit. 12*, 4 (Nov. 2019). doi:10.1145/3344383. 2

[ASGA17] ANCILLAO A., SAVASTANO B., GALLI M., ALBERTINI G.: Three dimensional motion capture applied to violin playing: A study on feasibility and characterization of the motor strategy. *Computer Methods and Programs in Biomedicine 149* (2017), 19–27. doi:https://doi.org/10.1016/j.cmpb.2017.07.005. 2

[AYA*23] ARISTIDOU A., YIANNAKIDIS A., ABERMAN K., COHEN-OR D., SHAMIR A., CHRYSANTHOU Y.: Rhythm is a dancer: Music-driven motion synthesis with global structure. *IEEE Transactions on Visualization and Computer Graphics 29*, 8 (8 2023), 3519–3534. doi:10.1109/TVCG.2022.3163676. 2

[Bar23] BARNETT J.: The ethical implications of generative audio models: A systematic literature review. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society* (2023), pp. 146–161. 13

[BBR*22] BITTNER R. M., BOSCH J. J., RUBINSTEIN D., MESEGUER-BROCAL G., EWERT S.: A lightweight instrument-agnostic model for polyphonic note transcription and multipitch estimation. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing* (Singapore, 2022), ICASSP'22. 2, 5

[BKD12] BENETOS E., KLAPURI A., DIXON S.: Score-informed transcription for automatic piano tutoring. In *2012 Proceedings of the 20th European Signal Processing Conference (EUSIPCO)* (2012), pp. 2153–2157. 2

[BPH*24] BROOKS T., PEEBLES B., HOLMES C., DEPUE W., GUO Y., JING L., SCHNURR D., TAYLOR J., LUHMAN T., LUHMAN E., NG C., WANG R., RAMESH A.: Video generation models as world simulators, 2024. Accessed: September: 2024. URL: https://openai.com/research/video-generation-models-as-world-simulators. 1

[BVGLH17] BAZZICA A., VAN GEMERT J., LIEM C. C., HANJALIC A.: Vision-based detection of acoustic timed events: a case study on clarinet note onsets. *arXiv preprint arXiv:1706.09556* (2017). 3

[BXP*22] BHATNAGAR B. L., XIE X., PETROV I., SMINCHISESCU C., THEOBALT C., PONS-MOLL G.: Behave: Dataset and method for tracking human object interactions. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (jun 2022), IEEE. 2

[CAB*24] CHHATRE K., ATHANASIOU N., BECHERINI G., PETERS C., BLACK M. J., BOLKART T., ET AL.: Emotional speech-driven 3d body animation via disentangled latent diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2024), pp. 1942–1953. 2

[CMU03] CMU GRAPHICS LAB: Carnegie-mellon motion capture (mocap) database. http://mocap.cs.cmu.edu/, 2003. Carnegie Mellon University. 1, 2

[Coc24] COCKOS: Reaper - digital audio workstation, 2024. Accessed: April 2024. URL: https://www.reaper.fm/index.php. 5

[Com24] COMMONS C.: Attribution-noncommercial–sharealike 4.0 international, 2024. Accessed: Jan 2025. URL: https://creativecommons.org/licenses/by-nc-sa/4.0/. 10

[EBD10] EMIYA V., BADEAU R., DAVID B.: Multipitch estimation of piano sounds using a new probabilistic spectral smoothness principle. *IEEE Transactions on Audio, Speech, and Language Processing 18*, 6 (2010), 1643–1654. doi:10.1109/TASL.2009.2038819. 3

[FDS*24] FAN Z., DAI P., SU Z., GAO X., LV Z., ZHANG J., DU T., WANG G., ZHANG Y.: Emhi: A multimodal egocentric human motion dataset with hmd and body-worn imus, 2024. arXiv:2408.17168. 1

[Foc24] FOCUSRITE: Scarlett - audio interface, 2024. Accessed: April 2024. URL: https://focusrite.com/scarlett. 5

[For24] FORTNITE: Fortnite festival season 4: Prepare for battle with metallica!, 2024. Accessed: September 2024. URL: https://www.fortnite.com/news/fortnite-festival-season-4-prepare-for-battle-with-metallica. 2

[FTT*23] FAN Z., TAHERI O., TZIONAS D., KOCABAS M., KAUFMANN M., BLACK M. J., HILLIGES O.: ARCTIC: A dataset for dexterous bimanual hand-object manipulation. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2023). 2

[GR06] GILLET O., RICHARD G.: Enst-drums: an extensive audio-visual database for drum signals processing. In *Proceedings of the 7th International Conference on Music Information Retrieval* (2006), ISMIR, pp. 156–159. 3

[GZW*20] GUO C., ZUO X., WANG S., ZOU S., SUN Q., DENG A., GONG M., CHENG L.: Action2motion: Conditioned generation of 3d human motions. In *Proceedings of the 28th ACM International Conference on Multimedia* (Oct. 2020), MM '20, ACM. doi:10.1145/3394171.3413635. 2

[Har23] HARRER S.: Attention is not all you need: the complicated case of ethically using large language models in healthcare and medicine. *EBioMedicine 90* (2023). 13

[HKS17] HOLDEN D., KOMURA T., SAITO J.: Phase-functioned neural networks for character control. *ACM Trans. Graph. 36*, 4 (July 2017). doi:10.1145/3072959.3073663. 2

[HMC*24] HUANG Y.-F., MORAN N., COLEMAN S., KELLY J., WEI S.-H., CHEN P.-Y., HUANG Y.-H., CHEN T.-P., KUO Y.-C., WEI Y.-C., LI C.-H., HUANG D.-Y., KAO H.-K., LIN T.-W., SU L.: Mosa: Music motion with semantic annotation dataset for cross-modal music processing. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* (2024), 1–14. doi:10.1109/TASLP.2024.3407529. 2, 3

[IPOS14] IONESCU C., PAPAVA D., OLARU V., SMINCHISESCU C.: Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence 36*, 7 (2014), 1325–1339. doi:10.1109/TPAMI.2013.248. 1, 2

[JQS*24] JIN Y., QIU Z., SHI Y., SUN S., WANG C., PAN D., ZHAO J., LIANG Z., WANG Y., LI X., YU F., YU T., DAI Q.: Audio matters too! enhancing markerless motion capture with audio signals for string performance capture. *ACM Trans. Graph. 43*, 4 (7 2024). doi:10.1145/3658235. 3

[JSS*24] JUNG S., SEOL Y., SEO K., NA H., KIM S., TAN V., NOH J.: Speed-aware audio-driven speech animation using adaptive windows. *ACM Trans. Graph.* (8 2024). doi:10.1145/3691341. 2

[JZL*24] JIANG N., ZHANG Z., LI H., MA X., WANG Z., CHEN Y., LIU T., ZHU Y., HUANG S.: Scaling up dynamic human-scene interaction modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2024), pp. 1737–1747. 2

[KAdlCCP*24] KYRIAKOU T., ALVAREZ DE LA CAMPA CRESPO M., PANAYIOTOU A., CHRYSANTHOU Y., CHARALAMBOUS P., ARISTIDOU A.: Virtual Instrument Performances (VIP): A comprehensive review. *Comput. Graph. Forum 43*, 2 (2024), 35–58. doi:10.1111:cgf.15065. 3, 6

[LAZ*22] LI P., ABERMAN K., ZHANG Z., HANOCKA R., SORKINE-HORNUNG O.: Ganimator: Neural motion synthesis from a single sequence. *ACM Transactions on Graphics (TOG) 41*, 4 (2022), 138. 2

[LDM*19] LEE G., DENG Z., MA S., SHIRATORI T., SRINIVASA S., SHEIKH Y.: Talking with hands 16.2m: A large-scale dataset of synchronized body-finger motion and audio for conversational motion analysis and synthesis. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)* (2019), pp. 763–772. doi:10.1109/ICCV.2019.00085. 2

[Leg20] LEGEND J.: John legend live - a night for "bigger love" presented by wave, 2020. URL: https://www.youtube.com/watch?v=eGy64l9Yuuw. 2

[LLD*19] LI B., LIU X., DINESH K., DUAN Z., SHARMA G.: Creating a multitrack classical music performance dataset for multimodal music analysis: Challenges, insights, and applications. *IEEE Transactions on Multimedia 21*, 2 (2019), 522–535. doi:10.1109/TMM.2018.2856090. 3

[LSP*20] LIU J., SHAHROUDY A., PEREZ M., WANG G., DUAN L.-Y., KOT A. C.: Ntu rgb+d 120: A large-scale benchmark for 3d human activity understanding. *IEEE Trans. Pattern Anal. Mach. Intell. 42*, 10 (Oct. 2020), 2684–2701. doi:10.1109/TPAMI.2019.2916873. 2

[LYRK21] LI R., YANG S., ROSS D. A., KANAZAWA A.: Ai choreographer: Music conditioned 3d dance generation with aist++. In *IEEE/CVF International Conference on Computer Vision* (2021), ICCV'21, pp. 13381–13392. doi:10.1109/ICCV48922.2021.01315. 2

[LZL*24a] LIANG H., ZHANG W., LI W., YU J., XU L.: Intergen: Diffusion-based multi-human motion generation under complex interactions. *International Journal of Computer Vision 132*, 9 (Mar. 2024), 3463–3483. doi:10.1007/s11263-024-02042-6. 2

[LZL*24b] LIN J., ZENG A., LU S., CAI Y., ZHANG R., WANG H., ZHANG L.: Motion-x: A large-scale 3d expressive whole-body human motion dataset, 2024. URL: https://arxiv.org/abs/2307.00818, arXiv:2307.00818. 3

[LZW*23] LIN J., ZENG A., WANG H., ZHANG L., LI Y.: One-stage 3d whole-body mesh recovery with component aware transformer. *CVPR* (2023). 3

[LZZ*23] LI R., ZHAO J., ZHANG Y., SU M., REN Z., ZHANG H., TANG Y., LI X.: Finedance: A fine-grained choreography dataset for 3d full body dance generation, 2023. URL: https://arxiv.org/abs/2212.03741, arXiv:2212.03741. 2

[Met22] META: Catch foo fighters in vr: Horizon venues concert to air february 13 after the big game, 2022. Accessed: September 2024. URL: https://www.meta.com/blog/quest/catch-foo-fighters-in-vr-horizon-venues-concert-to-air-february-13-after-the-big-game/. 2

[MGT*19] MAHMOOD N., GHORBANI N., TROJE N. F., PONS-MOLL G., BLACK M. J.: AMASS: Archive of motion capture as surface shapes. In *International Conference on Computer Vision* (Oct. 2019), pp. 5442–5451. 1, 2

[Mov23] MOVELLA: The story behind the virtual concert experience of twenty one pilots in the roblox metaverse using xsens motion capture technology., 2023. Accessed: September 2024. URL: https://www.movella.com/resources/cases/the-story-behind-the-virtual-concert-experience-of-twenty-one-pilots-in-the-roblox-metaverse-using-xsens-motion-capture-technology. 2

[MRPM14] MARCHINI M., RAMIREZ R., PAPIOTIS P., MAESTRE E.: The sense of ensemble: a machine learning approach to expressive performance modelling in string quartets. *Journal of New Music Research 43*, 3 (2014), 303–317. doi:10.1080/09298215.2014.922999. 3

[MSH20] MONTESINOS J. F., SLIZOVSKAIA O., HARO G.: Solos: A dataset for audio-visual music analysis. In *IEEE 22nd International Workshop on Multimedia Signal Processing* (2020), MMSP, pp. 1–6. doi:10.1109/MMSP48831.2020.9287124. 3

[MYH*24] MA L., YE Y., HONG F., GUZOV V., JIANG Y., POSTYENI R., PESQUEIRA L., GAMINO A., BAIYYA V., KIM H. J., BAILEY K., FOSAS D. S., LIU C. K., LIU Z., ENGEL J., NARDI R. D., NEWCOMBE R.: Nymeria: A massive collection of multimodal egocentric daily motion in the wild, 2024. arXiv:2406.09905. 1

[Ope22] OPENAI: Introducing chatgpt, 2022. Accessed: September: 2023. URL: https://openai.com/index/chatgpt/. 1

[PCAW16] PEREZ-CARRILLO A., ARCOS J.-L., WANDERLEY M.: Estimation of guitar fingering and plucking controls based on multimodal analysis of motion, audio and musical score. In *11th International Symposium on Music, Mind, and Embodiment* (2016), Springer, pp. 71–87. doi:10.1007/978-3-319-46282-0_5. 3

[PMA16] PLAPPERT M., MANDERY C., ASFOUR T.: The kit motion-language dataset. *Big Data 4*, 4 (Dec. 2016), 236–252. doi:10.1089/big.2016.0028. 2

[PMPCM14] PAPIOTIS P., MARCHINI M., PEREZ-CARRILLO A., MAESTRE E.: Measuring ensemble interdependence in a string quartet through analysis of multidimensional performance data. *Frontiers in Psychology 5* (2014). doi:10.3389/fpsyg.2014.00963. 3

[Rok23] ROKOKO: Actor profile measurements, 2023. Accessed: October 2024. URL: https://youtu.be/P_9P2hThExU. 7

[Rok24a] ROKOKO: Five coil pro features that will transform your mocap workflow, 2024. Accessed: April 2024. URL: https://www.rokoko.com/insights/five-coil-pro-features-that-will-transform-your-mocap-workflow. 4

[Rok24b] ROKOKO: Track body, finger and facial motions with full performance capture, 2024. Accessed: April 2024. URL: https://www.rokoko.com/products/full-performance-capture. 4

[RPG*21] RAMESH A., PAVLOV M., GOH G., GRAY S., VOSS C., RADFORD A., CHEN M., SUTSKEVER I.: Zero-shot text-to-image generation. In *Proceedings of the 38th International Conference on Machine Learning* (7 2021), Meila M., Zhang T., (Eds.), vol. 139 of *Proceedings of Machine Learning Research*, PMLR, pp. 8821–8831. 1

[RVM24] RONSSIN D., VASS T., MOREL P.: Neuralnote, 2024. Accessed: April 2024. URL: https://github.com/DamRsn/NeuralNote. 5

[SCTO17] SARASÚA A., CARAMIAUX B., TANAKA A., ORTIZ M.: Datasets for the analysis of expressive musical gestures. In *Proceedings of the 4th International Conference on Movement Computing* (New York, NY, USA, 2017), MOCO '17, ACM. doi:10.1145/3077981.3078032. 3

[SGLH*20] SLATER M., GONZALEZ-LIENCRES C., HAGGARD P., VINKERS C., GREGORY-CLARKE R., JELLEY S., WATSON Z., BREEN G., SCHWARZ R., STEPTOE W., ET AL.: The ethics of realism in virtual and augmented reality. *Frontiers in Virtual Reality 1* (2020), 1. 13

[TLL*23] TANG Y., LIU J., LIU A., YANG B., DAI W., RAO Y., LU J., ZHOU J., LI X.: Flag3d: A 3d fitness activity dataset with language instruction, 2023. URL: https://arxiv.org/abs/2212.04638, arXiv:2212.04638. 2

[TRG*23] TEVET G., RAAB S., GORDON B., SHAFIR Y., COHEN-OR D., HAIM BERMANO A.: Human motion diffusion model. In *The Eleventh International Conference on Learning Representations* (2023), ICLR'23. URL: https://openreview.net/forum?id=SJ1kSyO2jwu. 2

[VKV*17] VOLPE G., KOLYKHALOVA K., VOLTA E., GHISIO S., WADDELL G., ALBORNO P., PIANA S., CANEPA C., RAMIREZ-MELENDEZ R.: A multimodal corpus for technology-enhanced learning of violin playing. In *Proceedings of the 12th Biannual Conference on Italian SIGCHI Chapter* (New York, NY, USA, 2017), CHItaly '17, ACM. doi:10.1145/3125571.3125588. 3

[WMR*21] WEIDINGER L., MELLOR J., RAUH M., GRIFFIN C., UESATO J., HUANG P.-S., CHENG M., GLAESE M., BALLE B., KASIRZADEH A., ET AL.: Ethical and social risks of harm from language models, 2021. arXiv:2112.04359. 13

[XBP*18] XI Q., BITTNER R. M., PAUWELS J., YE X., BELLO J. P.: Guitarset: A dataset for guitar transcription. In *19th International Society for Music Information Retrieval Conference* (2018). URL: https://api.semanticscholar.org/CorpusID:53875945. 3

[XTNK24] XU R., TRAN V. A., NAYAR S. K., KRISHNAN G.: Dancecraft: A music-reactive real-time dance improv system. In *Proceedings of the 9th International Conference on Movement and Computing* (New York, NY, USA, 2024), MOCO '24, ACM. doi:10.1145/3658852.3659078. 2

[ZCP*22] ZHANG M., CAI Z., PAN L., HONG F., GUO X., YANG L., LIU Z.: Motiondiffuse: Text-driven human motion generation with diffusion model, 2022. arXiv:2208.15001. 2

[ZHL*24] ZHANG Y., HUANG D., LIU B., TANG S., LU Y., CHEN L., BAI L., CHU Q., YU N., OUYANG W.: Motiongpt: Finetuned llms are general-purpose motion generators, 2024. arXiv:2306.10900. 2

[ZZC*23] ZHANG J., ZHANG Y., CUN X., HUANG S., ZHANG Y., ZHAO H., LU H., SHEN X.: T2m-gpt: Generating human motion from textual descriptions with discrete representations, 2023. arXiv:2301.06052. 2

Appendix A: Performers biography

The biographies of the four professional musicians who participated in the recording sessions.

- One male guitarist with 20 years of experience in the field. He studied jazz composition at Berklee College of Music, graduating with honors with a Bachelor of Music, and later pursued

a Master of Fine Arts in jazz guitar, followed by a Doctor of Musical Arts in composition and performance at the California Institute of the Arts.

- One female pianist with 23 years of experience in classical piano performance. She completed her Bachelor of Music with honors in Piano Performance at the Royal Northern College of Music (RNCM), specializing in classical music. In her final year, she was awarded the Clifford Hartley Award for her achievements in the Advanced Piano Teaching course. She then completed a Postgraduate Certificate in Education (PGCE) in Specialist Instrumental Teaching at RNCM and Manchester Metropolitan University (MMU). Additionally, she holds a Master's degree in Music Education, further enhancing her contributions to the study.
- One male pianist with 20 years of experience in piano performance. He studied piano at Royal Holloway, University of London, specializing in composition, where he graduated with First Class Honours. He then continued his studies at the same university, earning a Master of Arts in Advanced Musical Studies, with a focus on composition.
- One male drummer with 29 years of experience in percussion. He studied at the Musicians Institute College of Contemporary Music in Hollywood, Los Angeles, California. During his time there, he was honored with the 'Outstanding Player of the Year' award and was featured in publications such as Modern Drummer magazine. He also studied classical percussion at the Arte Music Academy. His diverse background in both contemporary and classical percussion provided valuable contributions to the study.