Music Performance Hands-included-Motion Generation via dual domain loss with Audio Reconstruction

Hiroki Nishizawa[1] Seong Jong Yoo[2] Keitaro Tanaka[3] Shugo Yamaguchi[1]
Qi Feng[3] Masatoshi Hamanaka[4] Cornelia Fermüller[2] Shigeo Morishima[3]

[1]Waseda University [2]University of Maryland (UMD)
[3]WISE, Waseda Research Institute for Science and Engineering [4]RIKEN

Abstract

Generating a natural musical instrument performance motion from audio input is crucial for entertainment products, but it remains extremely challenging due to the complex mapping between audio and motion. Recent studies rely on costly information or instrument-dominant methods, which leads to difficulties in scaling datasets and exploring the variety of instrument motion generation. Furthermore, recent studies do not consider whether the generated motion actually plays the music or not. In this work, we propose a novel cyclic approach that optimizes the motion generator with audio feedback through the pre-trained audio generator. This approach enables softly guaranteeing the motion that actually plays the input music. Additionally, using the pre-trained audio generator, we propose a new motion metric called ARD (Audio Reconstruction Distance) that evaluates the acoustical accuracy of the motion. The proposed method achieved competitive results in spatial accuracy, but it was more acoustically correct than the baseline methods.

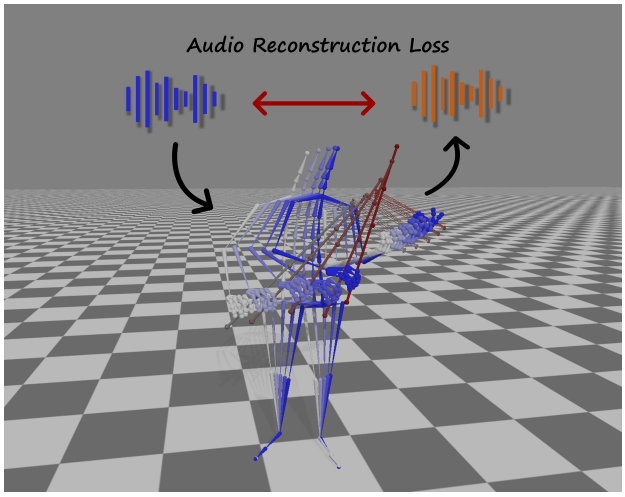

Figure 1. **Core idea of our cyclic method.** Musical instruments can be modeled as a complex function that maps human motion collisions into audio signals. Building on this notion, we introduce a cyclic regularization term that ensures the motion indeed produces the intended sounds.

1. Introduction

Instrumental performance is not merely the enjoyment of sound produced with precision; it is an art in which the performer conveys their interpretation of music through the movements of the body. This artistry is not confined to a single instrument but is shared across many, and through the interplay and harmonics of multiple instruments, it has been refined and elevated into an ever more profound form of expression.

According to the Virtual Instrument Performances (VIP) survey [5], instrumental performance motion generation is expected to find applications in film, animation, and virtual concerts; accordingly, the ability to automatically generate natural performance motions that are synchronized with audio is indispensable.

With advances in deep learning, a line of research has explored generating instrumental performance motion from audio using deep neural networks [2, 4, 9]. However, because of simplified input–output settings and model designs, the resulting motions have not reached artistic fidelity, and generating full-body motion—including the hands remains difficult. To address this, Nishizawa et al. [7] adopted an LSTM-based multistage approach that employs performance information as an intermediate representation, thereby enabling, for the first time, the generation of natural full-body performance motion. Building on progress in generative modeling, Qiu et al[8]. further made highly natural performance motion possible by incorporating a diffusion model together with instrument-collision detection

into the loss function. Nevertheless, these methods exhibit strong instrument dependence and require costly annotation and model-specific engineering, which limits dataset scaling and hinders generalization across instrument types.

To this end, we propose a new cyclic method for instrumental performance motion generation with low instrument dependence. Specifically, we use a pretrained audio reconstruction model to convert the output motion of the motion-generation model back into audio and compute losses in both the motion and audio domains. This not only enables efficient training of the motion-generation model but also facilitates transfer to a wide range of instruments. Moreover, leveraging this pretrained audio reconstruction model, we introduce ARD, a metric that quantifies how acoustically correct the generated motion is.

In summary, our main contributions are twofold:

- A new cyclic method for performance motion generation enables efficient learning while keeping instrument dependence low.
- A new evaluation metric (ARD) assesses whether the generated motion is acoustically correct.

2. Dataset

Motion data of instrumental performance is an extremely valuable resource in this field, and numerous datasets have been proposed. The majority of these are captured using engineering-style marker-based systems. While there exist datasets covering various instruments and large-scale collections, most omit the hands due to the complexity of motion capture and the difficulty of performing under such conditions. Nishizawa et al. proposed a violin performance motion dataset that includes the hands, but it is limited to three hours in duration and does not record instrument positions.

To reduce the burden on performers, markerless motion capture systems have been proposed as an alternative [3]. However, because of their markerless nature, their accuracy has been shown—through subjective evaluation by professional violinists—to be inferior to that of marker-based datasets[7].

Therefore, we employed a marker-based motion capture system designed to minimize the physical load on the hands, as illustrated in Figure 2. Furthermore, by attaching markers to the instrument as shown in Figs 3 and 4, we successfully recorded a motion dataset encompassing the full body, including the hands, as well as the bow and violin. In total, we recorded approximately 6 hours of performance data from 81 pieces by 6 professional violinists.

3. Method

We propose a novel cyclic method that optimizes the motion generation model by both the motion of the domains and audio. Our method consists of two modules, the audio generator and the motion diffusion model, which are described in 3.1 and 3.2. We combined the two methods into one, as described in 3.3.

Figure 2. **Hand markers.** We used physically friendly markers that are very light and have minimal effect on playing a piece of music.

Figure 3. **Bow markers.** We set four markers in total to capture the bow location.

Figure 4. **Violin markers.** We used a MIDI violin and set five markers on it to capture its location.

3.1. Audio Generator: $\mathcal{A}$

To enable cyclic training, we train an audio generator $\mathcal{A}$ that reconstructs audio from the motion input $\mathbf{x}_0$. As illustrated in Fig. 3.1, $\mathcal{A}$ is implemented with a transformer-based encoder–decoder architecture.

For efficiency, we leverage the prior knowledge that the most relevant body parts for audio generation are the hands and arms; thus, only these joints are used as the body input. Moreover, the positions of the violin and bow are incorporated as auxiliary information to enhance audio gen-

eration. The audio generator outputs a music-intrinsic representation in the form of MIDI, which consists of four attributes: velocity (note intensity), onset (note activation timing), pitch (discrete note value).

MIDI Representations and Conditions. In summary, the inputs to the audio generator comprise the joint locations of the hands and arms together with the violin and bow positions. The output is the MIDI representation, which consists of velocity, pitch, and onset.

3.2. Motion Generator: $\mathcal{M}$

Following recent advancements in diffusion models for motion synthesis[8], we designed a DDPM-based method as our baseline model. The diffusion transformers with long skip connections are a common network architecture for stable and fast training [1]. The overview of $\mathcal{M}$ is illustrated in Figure 6. We designed a two-linear and multi-head attention layer as an initial projection, a five-layer transformer encoder, a four-layer transformer decoder with long skip connections, and one linear block as a last projection.

The forward trajectory is described as Eq 1, $\mathcal{N}$ denotes a Gaussian distribution.

$$q(\mathbf{x}_t \mid \mathbf{x}_{t-1}) = \mathcal{N}\Big(\mathbf{x}_t;\ \sqrt{\alpha_t}\,\mathbf{x}_{t-1},\ (1-\alpha_t)\mathbf{I}\Big) \quad (1)$$

$$\alpha_t = 1 - \beta_t, \qquad \bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s, \quad t = 1, \ldots, T \quad (2)$$

Unlike conventional DDPMs, which estimate the one-step posterior $p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t)$, our method directly estimates $\mathbf{x}_0$ from each time step in order to compute the audio reconstruction loss. Accordingly, the reverse diffusion process estimated by the model is defined in Eqs 3 and 4.

$$p_\theta(\mathbf{x}_0 \mid \mathbf{x}_t) = \mathcal{N}\Big(\mathbf{x}_0;\ \tilde{\boldsymbol{\mu}}_\theta(\mathbf{x}_t, t),\ \tilde{\boldsymbol{\Sigma}}_\theta(t)\Big), \quad (3)$$

$$\tilde{\boldsymbol{\mu}}_\theta(\mathbf{x}_t, t) = \frac{1}{\sqrt{\bar{\alpha}_t}}\Big(\mathbf{x}_t - \sqrt{1-\bar{\alpha}_t}\,\boldsymbol{\varepsilon}_\theta(\mathbf{x}_t, t)\Big), \quad (4)$$

In practice, diffusion models are trained by minimizing a noise-prediction mean squared error (MSE), which is a simplified form of the ELBO objective[6] defined in eq 5.

$$\mathcal{L}_{\text{pose}} = \big\|\mathbf{x}_0 - \hat{\mathbf{x}}_0\big\|_2 \quad (5)$$

Motion Representations and Conditions. The motion $\mathbf{x}_0$ consists of 73 joint locations (24 for each hand and 25 for the remaining body parts), along with 6-DoF (*rotation and translation*) 4×3 matrices for both the violin and the bow. Formally, the motion tensor is represented as $\mathbf{x}_0 \in \mathbb{R}^{F \times (J+4) \times 3}$, where F denotes the number of frames and $J = 73$ is the number of body joints. All motion and bow representations are defined in the violin-local coordinate system, and the 6-DoF pose of the violin is used to transform these into the global coordinate system. Details of this localization procedure are described in sec 4.1.

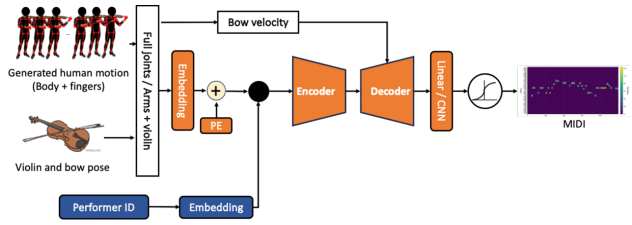

Figure 5. **Overview of the audio generator.** The model is designed as a transformer-based diffusion framework with long skip connections to accelerate convergence and enable stable training.

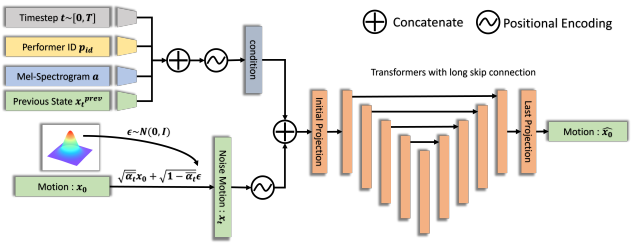

Figure 6. **Model architecture.** Detailed design of the transformer-based diffusion model, highlighting the long skip connections that improve convergence speed and training stability.

The condition input to $\mathcal{M}$ comprises four components: the audio feature $\mathbf{a}$, the performer ID ($\mathbf{p}_{\text{id}}$), the diffusion timestep t, and the previous four frames of motion $\mathbf{x}^{\text{prev}}$. Specifically, we use a 128-dimensional mel-spectrogram as the audio feature, a 6-dimensional one-hot vector for the performer ID ($\mathbf{p}_{\text{id}}$), a scalar value for the timestep t, and the same representation format for the previous motion state $\mathbf{x}^{\text{prev}}$.

3.3. A Combination of $\mathcal{A}$ and $\mathcal{M}$

We integrate the two modules by optimizing the motion diffusion model with the joint loss $\mathcal{L}_{\text{dual}}$, which combines the pose reconstruction loss and the audio reconstruction loss. The complete objective is defined in Eqs. 6 and 7.

$$\mathcal{L}_{\text{ARD}} = \big\|\mathbf{MIDI} - \mathcal{A}(\hat{x}_0)\big\|_2, \quad (6)$$

$$\mathcal{L}_{\text{dual}} = \mathcal{L}_{\text{pose}} + \lambda_{\text{ARD}}\,\mathcal{L}_{\text{ARD}}, \quad (7)$$

4. Experiments

4.1. Experiment Setting

We trained the Motion Generator for 300k iterations with 260 batches for 8-second motion sequences (60 fps). In addition, we trained for another 100k iterations with and with-

out the audio reconstruction loss. We denoised 1k steps for each model. Both methods estimate violin-localized coordinated motion.

Specifically, we employ the 6-DoF pose of the violin, consisting of a rotation matrix $\mathbf{R}_t \in \mathbb{R}^{3\times3}$ and a translation vector $\mathbf{t}_t \in \mathbb{R}^3$, to normalize the violin position at the reference frame to the origin. Using this reference, all subsequent frames are transformed into the violin-local coordinate system by applying the same rotation and translation:

$$\tilde{\mathbf{x}}_t = \mathbf{R}_t^\top (\mathbf{x}_t - \mathbf{t}_t), \quad t = 1, \ldots, F, \tag{8}$$

where $\mathbf{x}_t \in \mathbb{R}^3$ denotes the global coordinate of a joint at frame t, and $\tilde{\mathbf{x}}_t$ is its representation in the violin-local coordinate system. This procedure prevents violin position estimation errors from propagating into the audio generation model, thereby providing more effective feedback.

4.2. Evaluation Metrics

We define the naturalness of the generated motion in terms of four aspects: (i) accuracy of joint positions, (ii) accuracy of joint trajectories, (iii) smoothness of motion, and (iv) audio–motion consistency. Each aspect is quantitatively evaluated using the corresponding metrics, namely L_1, DTW, Jerk, and ARD, as formulated in Eqs. 9 - 12.

L1 Loss.

$$L_1 = \frac{1}{F}\sum_{i=1}^{F} \left\| x_i^{\text{gt}} - x_i^{\text{pred}} \right\|_1 \tag{9}$$

Dynamic Time Warping (DTW).

$$\text{DTW} = \min \frac{1}{F} \sum_{(i,j)\in\mathcal{F}} \left\| x_i^{\text{gt}} - x_j^{\text{pred}} \right\|_1 \tag{10}$$

Jerk.

$$\text{Jerk} = \frac{1}{F}\sum_{i=1}^{F} \left\| \frac{\partial^3 x^{\text{pred}}}{\partial i^3} \right\|_1 \tag{11}$$

Audio Reconstruction Loss (ARD).

$$\text{ARD} = \frac{1}{F}\sum_{i=1}^{F} \left\| \text{MIDI}_i - \mathcal{A}(\hat{x}_i) \right\|_2 \tag{12}$$

4.3. Quantitative Results

Table 1 shows the quantitative results of the baseline Motion Generator and our proposed method with the cyclic audio reconstruction loss. The results demonstrate that our method achieves comparable performance to the baseline in terms of spatial accuracy (L1 and DTW), while achieving higher scores in audio consistency. This confirms the effectiveness of our cyclic approach in generating acoustically coherent and better synchronized audio–motion pairs.

Table 1. **Quantitative evaluation on motion naturalness.** Comparison of different methods using L1 distance, Dynamic Time Warping (DTW), Jerk, and ARD metrics. Lower values indicate better performance.

Method	L1	DTW	Jerk	ARD
Baseline: $\mathcal{M}$	0.0238	**257**	0.0269	0.0138
+ w/ $\mathcal{L}_{\text{ARD}}$	**0.0238**	258	**0.0265**	**0.0133**
Ground Truth	–	–	0.000193	0.0104

However, our approach does not yet reach state-of-the-art performance on spatial metrics. The primary reason lies in the limitations of the Audio Generator. The proposed Audio Generator is trained on MIDI representations, and due to the discretized nature of this format, it is challenging to map continuous human motion into such a space accurately. Furthermore, our model does not use contact information, which reduces the robustness of the Audio Generator.

5. Conclusion and Future Work

We introduced a cyclic framework for audio-conditioned performance motion generation that couples a DDPM-based motion generator $\mathcal{M}$ with a pretrained audio generator $\mathcal{A}$. By reconstructing audio from motion and optimizing losses in both the motion and audio domains, the proposed scheme softly enforces that the generated motion actually plays the input music. In addition, we proposed *Audio Reconstruction Distance* (ARD) as a metric to quantify the acoustic correctness of motion via $\mathcal{A}$. Quantitative results (Table 1) show that our method achieves performance comparable to a strong baseline in spatial accuracy (L1, DTW) while delivering superior audio consistency, yielding acoustically coherent and better synchronized audio–motion pairs. These gains are obtained without relying on instrument-specific heuristics, highlighting the promise of the cyclic objective for scalable, instrument-agnostic learning. Our approach does not yet reach SoTA on spatial metrics, primarily due to limitations of the audio generator trained on discretized MIDI representations and the absence of explicit contact cues.

In future work, we plan to (i) incorporate contact and interaction signals (e.g., bow–string and hand–instrument events), (ii) employ richer or continuous audio targets (e.g., mel or differentiable waveform renderers) to reduce the discretization gap, (iii) explore joint or alternating training of $\mathcal{A}$ and $\mathcal{M}$ with stability safeguards, and (iv) extend evaluations beyond violin to multiple instruments and ensemble settings, including perceptual studies. We believe the cyclic paradigm and ARD provide general tools for advancing audio-synchronized, physically plausible performance motion generation.

References

[1] Guanjie Chen, Xinyu Zhao, Yucheng Zhou, Xiaoye Qu, Tianlong Chen, and Yu Cheng. Towards stabilized and efficient diffusion transformers through long-skip-connections with spectral constraints. *arXiv preprint arXiv:2411.17616*, 2024. 3

[2] Jiali Chen, Changjie Fan, Zhimeng Zhang, Gongzheng Li, Zeng Zhao, Zhigang Deng, and Yu Ding. A music-driven deep generative adversarial model for guzheng playing animation. *IEEE Transactions on Visualization and Computer Graphics*, 29(2):1400–1414, 2021. 1

[3] Yitong Jin, Zhiping Qiu, Yi Shi, Shuangpeng Sun, Chongwu Wang, Donghao Pan, Jiachen Zhao, Zhenghao Liang, Yuan Wang, Xiaobing Li, et al. Audio matters too! enhancing markerless motion capture with audio signals for string performance capture. *ACM Transactions on Graphics (TOG)*, 43 (4):1–10, 2024. 2

[4] Hsuan-Kai Kao and Li Su. Temporally guided music-to-body-movement generation. In *Proceedings of the 28th ACM international conference on multimedia*, pages 147–155, 2020. 1

[5] Theodoros Kyriakou, M Álvarez de la Campa Crespo, Andreas Panayiotou, Yiorgos Chrysanthou, Panayiotis Charalambous, and Andreas Aristidou. Virtual instrument performances (vip): A comprehensive review. In *Computer Graphics Forum*, page e15065. Wiley Online Library, 2024. 1

[6] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International conference on machine learning*, pages 8162–8171. PMLR, 2021. 3

[7] Hiroki Nishizawa, Keitaro Tanaka, Asuka Hirata, Shugo Yamaguchi, Qi Feng, Masatoshi Hamanaka, and Shigeo Morishima. Syncviolinist: Music-oriented violin motion generation based on bowing and fingering. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 5419–5428. IEEE, 2025. 1, 2

[8] Zhiping Qiu, Yitong Jin, Yuan Wang, Yi Shi, Chao Tan, Chongwu Wang, Xiaobing Li, Feng Yu, Tao Yu, and Qionghai Dai. Elgar: Expressive cello performance motion generation for audio rendition. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, pages 1–9, 2025. 1, 3

[9] Eli Shlizerman, Lucio Dery, Hayden Schoen, and Ira Kemelmacher-Shlizerman. Audio to body dynamics. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7574–7583, 2018. 1