From Improvised Movement to Musical Improvisation - Using Machine-Learning to Create Personalized Instruments for Dancers

Daniel Bisig*
daniel.bisig@zhdk.ch
Zurich University of the Arts
Zurich, Switzerland

Alexander Okupnik*
alexander.okupnik@uni.li
University of Liechtenstein
Vaduz, Liechtenstein
Zurich University of the Arts
Zurich, Switzerland

Johannes Schneider
johannes.schneider@uni.li
University of Liechtenstein
Vaduz, Liechtenstein

Diane Gemsch
dianegemsch@hotmail.com
Freelance
Zurich, Switzerland

Eleni Mylona
mylonaeleni@hotmail.com
Freelance
Zurich, Switzerland

Tim Winkler
tim_winkler@gmx.de
Freelance
Arnstadt, Germany

Abstract

The paper presents the development and preliminary evaluation of a machine-learning-based digital instrument that translates a dancer's bodily movements into music. It is trained on motion-capture and audio recordings of professional dancers improvising solo to music, thereby learning cross-modal correspondences between movement and sound. In performance, a dancer can then use the instrument as a highly personalized tool for generating music through bodily gestures. A transdisciplinary team of machine-learning researchers and dancers leads the development and evaluation, following a practice-led approach aligned with the dancers' artistic interests. This includes selecting the movement and musical material for training and testing, assessing the instrument's creative usability, and integrating it into rehearsals and the creation of new performance works.

Keywords

Generative Machine Learning, Movement to Music Translation, Dance and Technology, Practice-Led Approach

1 Introduction

In contemporary media-augmented dance performance, dancers increasingly act not only as interpreters of music but as active agents in its generation. This reciprocal relationship, in which dancers both shape and respond to the sonic environment, reframes movement and music as interdependent creative forces and addresses three concerns: reconceptualising the dancer–music relationship as dialogical rather than hierarchical, extending the dancer's presence beyond the physical body, and exploring translations of embodied expression into alternative modalities.

Informed by these concerns, the present work extends an earlier machine-learning (ML) system, *RAMFEM* (Raw Music From Free Movements) [4], which translates dancer movements into music. Designed as a bidirectional translator, *RAMFEM* captures the interpretive choices dancers make when improvising to music and then reverses this process to generate sound directly from movement. The new iteration, *PEMIDA* (PErsonalised Music Instrument for DAncers), develops this concept both artistically and technically: through a practice-led collaboration with multiple dancers, it integrates artistic considerations into system design, dataset curation, and performance evaluation, while introducing real-time inference, improved audio quality, and a more expressive ML architecture. All source code for *PEMIDA* is publicly available in a dedicated GitHub repository[1].

2 Background

This work builds on artistic and academic explorations of translating dancers' movements into music and relates to recent ML-based approaches to movement sonification. These include systems that map movement features to sound control parameters and systems that directly synthesise audio, enabled by recent advances in neural audio. The research adopts a practice-led methodology centred on dancers' active involvement, in line with artist-centred design in HCI.

2.1 Movement Sonification

This section reviews movement-to-music systems in dance contexts, grouped by technical approach: rule-based systems without ML, ML systems that map movement features to sound controls, and ML systems that directly synthesize audio from movement data.

2.1.1 Rule-based Approaches. Rule-based systems for controlling music through body movements have a long history. A prominent example is *Very Nervous System*, David Rokeby's computer-vision-based motion tracking system, which analyzes dancer movement from video and translates it into real-time musical output [28][35].

Since then, numerous movement sonification systems for dance have appeared. A real-time system maps body posture balance metrics to sound, encouraging dancers to explore unfamiliar movements [6]. Studies on conveying the movement quality "fluidity" through sound produced design guidelines linking specific sonic features to expressive motion [13]. Participatory work with dancers identified four key challenges: balancing automation

*All authors contributed equally to this research.

NIME '26, June 23–26, 2026, London, UK

[1] *PEMIDA* source code

and control, managing complex mappings intuitively, reconciling artistic and technical goals, and balancing understanding with enjoyment [21]. An interactive model-based sonification tool examined how real-time acoustic feedback supports sensorimotor learning via mappings between movement features and physical synthesis parameters [22]. A recent project treating motion-capture as machine-readable notation mapped dancer movement to sound synthesis and proposed ML to bridge discrete MIDI and continuous synthesis control [24].

2.1.2 ML-based Parameter Mapping. Several projects use ML to create adaptive mappings between bodily gestures and sound-engine parameters. Initially introduced when models could not generate audio directly, this strategy remains central in interactive ML frameworks [10] because it supports exploratory, iterative creative workflows.

One line of research targets intuitive control of high–dimensional synthesis spaces: a real-time system uses a multi-layer-perceptron to map hand-drawn gestures to modal synthesis parameters [18]. Another emphasizes personalized gesture–sound relationships, combining Gaussian-Mixture and Hidden-Markov-Regression to learn user-specific mappings from demonstrations, then driving descriptor-based granular synthesis via direct, temporal, or metaphorical mappings [12]. Co-creative performance systems integrate sequence prediction with adaptive mapping: *GestureRNN* predicts gestural trajectories from a pressure-sensitive interface, enabling collaborative musical phrase generation [16].

2.1.3 ML-based Motion to Audio Translation. Advances in ML have enabled end-to-end motion-to-audio systems that map raw sensor data directly to sound, building on neural audio synthesis with deep networks generating audio. Two main approaches exist: (1) conventional neural networks trained for audio synthesis, and (2) architectures that integrate differentiable digital-signal-processing (DDSP) components.

RAVE exemplifies the first approach [5], using a variational autoencoder with a filter-bank–plus–convolution encoder, a decoder with dilated residual stacks, and a multiscale discriminator; it combines representation learning with adversarial fine-tuning and singular value decomposition for compact latent control, achieving high-quality 48 kHz audio in faster-than–real-time.

The second approach, *DDSP* [9], combines neural networks with differentiable DSP modules such as harmonic additive and filtered-noise synthesis and trainable reverberation, whose interpretable components and strong inductive bias yield compact, data-efficient models.

The present work follows the first approach, extending *RAMFEM* [4], an earlier end-to-end model translating dance motion into raw audio via an adversarial autoencoder for waveform synthesis and a recurrent sequence-to-sequence transducer mapping motion to audio encodings, trained on recordings of a professional dancer; while sharing this project's goal of personalised, motion-responsive sound, *RAMFEM* produced lower audio quality and was unsuitable for real-time use.

Since *RAMFEM*, several works have appeared. This includes mapping low-dimensional performance parameters into *RAVE*'s latent space for creative control [34], genre-specific dance-music generation with a motion- and genre-conditioned latent diffusion model [31], embodied interaction with *RAVE* via several mapping strategies [26], real-time movement sonification by aligning body-pose and *RAVE* latent spaces [25], and dual- VAE–plus–GAN translation of motion and audio latents into new, fluid motion-controlled sound [33].

2.2 Artist-Centred Design

Beyond advances in model architecture, performance, and audio quality, this project also differs from *RAMFEM* through a practice-led, interdisciplinary methodology informed by professional dancers' expertise and interests. This perspective aligns with views that treat creative practitioners' active involvement as central to developing dance-related human–computer interaction tools [39]. The projects below exemplify cases where dancers substantially shaped both technical and artistic outcomes in motion-to-music translation.

Bevilacqua et al. present an interactive motion-capture system generating synchronized video and sound [3], showing that expressive gesture–sound relationships depend on mapping choices such as linking movement to timbre, introducing slight timing offsets, and maintaining perceptible audiovisual connections. Landry and Jeon apply participatory design to real-time movement sonification [21], where dancers negotiate the control—automation balance and influence mappings, yielding guidelines on agency, discoverability, and artistic immersion. Frid et al. examine how the movement quality *Fluidity* can be expressed sonically using dancer-generated motion data [13], identifying correspondences between continuous, low-register sounds and fluid motion, and discontinuous, high-frequency sounds for non-fluid motion. Bergsland's project on phrase boundaries in interactive dance sonification [2] uses iterative technical development and in-studio exploration to produce mappings that render phrase edges perceptually clear and artistically engaging. Finally, Nabi et al. investigate embodied navigation of neural audio synthesis latent spaces in live performance [26], with a dancer co-designing interaction strategies and reflecting on controllability, expressivity, and embodied understanding, positioning embodied exploration as a key design paradigm.

3 Artist Involvement

The present work is part of an ongoing collaboration between two ML researchers and three professional dancers—two established and one emerging—each of whom places sound at the centre of their artistic practice. Although they had no prior experience with interactive sound systems, all were keen to explore new ways of controlling sound through movement.

The dancers are referred to as Diane Gemsch (Dancer1), Eleni Mylona (Dancer2), and Tim Winkler (Dancer3). Their artistic profiles and relationships to music are briefly outlined below.

Dancer1 combines performance, choreography, and yoga-based mindful movement. She uses music and sound as catalysts for embodied awareness, emotional resonance, and relational exploration, examining how rhythm, silence, and musical texture evoke specific physical and spatial states.

Dancer2 bridges contemporary dance, performance, community practice, and artistic research, with a focus on language, feminist perspectives, and collective creation. She treats music and movement as interdependent, using sound to shape choreographic structures through live performance, original compositions, and curated playlists of emotional and political depth.

Dancer3 centres his practice on improvisation and instant composition, exploring what differentiates "dance" from "movement." Drawing on *Axis Syllabus* principles [11], he investigates biomechanical efficiency and structured improvisation, viewing music as an abstract yet accessible form and using dance to deepen its perception for audiences.

3.1 Improvisation and Recording Sessions

As training material for *PEMIDA*, movement and music were recorded for each dancer's solo improvisation, with each dancer selecting material from their own repertoire. Dancer1 chose excerpts from *The Anthropic Landscape – Life Beyond the Surface*[2], which intertwines her father's memory with choreographic and sonic elements, using a voice-recording of her father during one of his final moments of mental vitality. Dancer2 drew from the *Katines community performances*[3], a collective project on emancipation in which each woman chose a personally empowering song, yielding a playlist spanning Greek folk, Balkan dances, hip-hop, punk, techno, and Latin American genres. Dancer3 selected music by a composer he had previously collaborated with, valuing its distinctive glitch aesthetic; his movement and music came from two different works, and he used improvisation to explore their interplay.

Each dancer completed three solo improvisation sessions (5–15 minutes) with movement and sound recorded under three foci with deliberately open instructions: (1) free improvisation, (2) attention to temporal relationships, and (3) attention to timbral qualities.

For recording, dancers wore an Xsens Link suit[4]. To preserve real-time accuracy, motion data were streamed using two custom programs—one converting the native Xsens protocol to Open Sound Control (OSC) [36][5], the other recording timestamped OSC messages[6]. Data were captured at 240 Hz, downsampled to 30 Hz for OSC transmission, and sessions were video recorded and used for synchronising motion and audio material.

4 Model Architectures, Data Processing, and Training

PEMIDA consists of 3 main components (see figure 1):

A pre-trained neural vocoder, *Vocos* [29], converts audio waveforms to mel-spectrograms and vice versa. A custom *β-Variational Autoencoder* (*β*-VAE) [17] learns latent audio representations by auto-encoding mel-spectrograms[7]. A custom *Transformer* [32] takes sequences of dance poses and audio latents and predicts the next step of the audio-latent sequence[8].

4.1 Vocos

Vocos is a neural vocoder that predicts complex Short-Time Fourier Transform (STFT) coefficients, reconstructing audio via inverse FFT. This frequency-domain design offers computational efficiency and perceptual quality improvements over time-domain vocoders. For *PEMIDA*, we use a publicly available pre-trained *Vocos* checkpoint[9] operating at 48 kHz.

4.2 β-Variational Autoencoder

The custom VAE developed for *PEMIDA* uses a lightweight convolutional encoder–decoder architecture (Figure 2). The encoder comprises 2D convolutional layers with separate linear heads outputting latent mean μ and standard deviation σ, while the decoder mirrors this structure with a fully connected layer followed by transposed convolutions. In the chosen configuration,

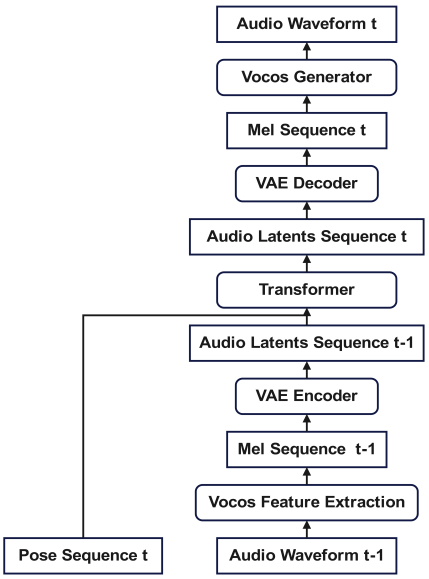

Figure 1: *PEMIDA*: Architectural components. Rectangles denote data, rounded rectangles denote ML models, and arrows denote information flow.

the encoder maps 8 mel-spectrogram frames (128 mel bins each) to a 32-dimensional latent vector, and the decoder reconstructs 8 frames from this code.

This compact design is preferred over higher-capacity models such as *RAVE* because it trains quickly (around one hour on an NVIDIA RTX 4090) on only minutes of audio, enabling fast experimentation and evaluation cycles.

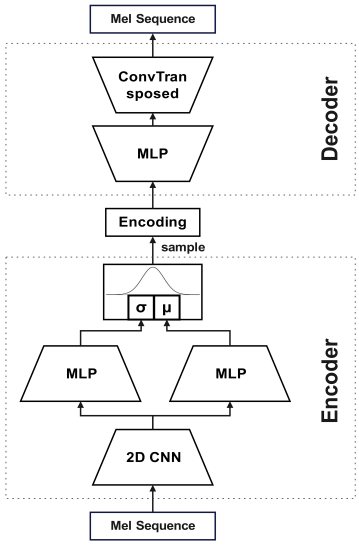

Figure 2: VAE architecture. Trapezoids denote layers that reduce dimensionality, inverted trapezoids those that increase it; rectangles indicate data representations, and arrows indicate data flow.

[2]Antropic-Landscape
[3]Katines-Community-Performances
[4]XSens Link Suit
[5]XSens2Osc Source Code
[6]SensorRecorder Source Code
[7]VAE Source Code
[8]Motion2Audio-Transformer Source Code
[9]kittn/vocos-mel-48khz-alpha1

4.3 Transformer

The custom Transformer (Figure 3) follows the standard encoder–decoder architecture of Vaswani et al. [32].

The encoder processes 60 motion-capture frames (2 seconds at 30 Hz), each consisting of 23 relative joint rotations in quaternion form. The decoder, conditioned on encoder outputs through cross-attention, receives 47 preceding normalized audio latent vectors and predicts the next latent vector, thus extending the audio sequence by one step.

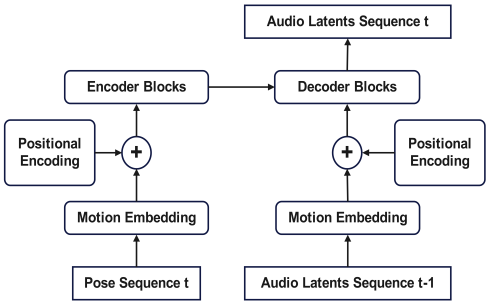

Figure 3: Transformer architecture. Rectangles denote data representations, rounded rectangles neural components, and arrows information flow.

4.4 Data Processing

4.4.1 Data Processing for VAE Training. The data processing procedure for the VAE segments continuous 48 kHz mono audio waveforms using a sliding window and transforms these segments into reshaped Mel spectrograms. The waveforms are divided into excerpts of 65,280 samples with a hop size of 960 samples. Each excerpt is passed through a *Vocos* feature extractor to obtain a Mel spectrogram with 128 Mel filter bands over 256 temporal frames. Since the Autoencoder operates on temporal contexts of 8 Mel frames, the spectrogram tensor is permuted and reshaped so that part of the temporal dimension is folded into the batch dimension, matching the model's input requirements.

4.4.2 Data Processing for Transformer Training. The data processing pipeline for the Transformer enforces temporal alignment between motion and audio, performs audio encoding, and applies normalization. Motion is represented as frame sequences, each frame encoding a pose of 23 joints via quaternion-based local rotations. Sequences are truncated to the frame range of the corresponding audio, the root joint's local rotation is set to the identity quaternion to remove global orientation, and each pose dimension is normalized using its mean and standard deviation over the full motion dataset.

Temporally aligned motion–audio excerpts for autoregressive training are generated with two nested loops. The outer loop applies a sliding window whose length equals the sum of the motion context and autoregressive target (64 context frames and 10 target frames). For each window, the inner loop iterates over the target motion frames: for the context and for each target frame, a temporally aligned audio excerpt is extracted and converted to a Mel spectrogram using *Vocos*, permuted to match the autoencoder's input format, and encoded into a low-dimensional latent vector. These latents are then normalized per dimension using statistics from the full audio dataset. Each training item thus contains a normalized input motion context, a normalized target motion sequence, a normalized input sequence of audio latents, and a stack of normalized target audio-latent sequences, one per target frame.

4.5 Training

Three datasets were created, each containing motion and audio from one dancer's improvisation session. For this study, only sessions focusing on temporal relationships between movement and sound were used to facilitate learning temporal correlations; remaining sessions will be included in future work.

Training proceeds in two stages with *Vocos* parameters fixed throughout: (1) the VAE is trained independently; (2) its weights are frozen and the Transformer is trained.

4.5.1 VAE Training. Audio is segmented into 65,280-sample excerpts (256 mel frames) for stable *Vocos* reconstruction, and spectrograms are grouped into 8-frame blocks matching the autoencoder's context window. Two configurations differ by maximum β (0.1 and 1.0) and whether the decoder is fine-tuned afterward, probing different trade-offs between reconstruction fidelity and latent-space regularization. Following Fu et al. [14], a cyclical β schedule is used. Joint encoder–decoder training runs for 400 epochs with step-decayed learning rate, followed by an optional 200-epoch decoder-only phase. The loss combines (Kullback–Leibler) KL divergence [20], mean squared error (MSE) on mel-spectrograms, and a multiresolution STFT loss [30] on reconstructed audio.

4.5.2 Transformer Training. For Transformer training, motion data are represented as normalized pose sequences of relative joint rotations in quaternion form and segmented into 70-frame overlapping windows; the first 60 frames form the encoder input and the remaining 10 frames define the autoregressive prediction horizon. During training, the decoder is first conditioned on ground-truth latents, then autoregressively on its own outputs to improve stability [1].

Three configurations are tested, differing by (i) the VAE version ($\beta = 0.1$ vs. $\beta = 1.0$ with decoder fine-tuning) and (ii) a joint-dropout scheme that zeros all but four randomly selected joint rotations to improve robustness to unseen joint configurations and enable control with a subset of joints. All configurations are trained for 200 epochs with step-decayed learning rates. The loss combines (1) MSE on audio latents, (2) MSE on decoded mel-spectrograms, and (3) a multiresolution STFT loss for waveform-level perceptual fidelity.

5 Evaluation Methods

Models were evaluated using quantitative, qualitative, and artistic approaches. Quantitative evaluation measures reconstruction fidelity and distributional match; qualitative evaluation measures audible differences and relies on critical listening; artistic evaluation uses dancers' feedback after testing the models through improvisation.

The evaluations were conducted with models from three training regimes: (i) an VAE with maximum $\beta = 0.1$ and a Transformer without and one with (ii) joint-dropout; and (iii) an VAE with maximum $\beta = 1.0$ plus decoder-only fine-tuning and a Transformer without joint-dropout.

5.1 Quantiative Evaluation

To quantitatively evaluate all models the following metrics are employed. Reconstruction fidelity is assessed via spectral features

using multi-resolution STFT loss (ms-STFT) and multi-resolution log-Mel loss (ms-log-Mel) as metrics. Linear-magnitude STFTs are compared by ms-STFT to capture fine detail and transients, while ms-log-Mel uses log-Mel spectrograms to reflect perceptual aspects such as timbre and spectral envelope. Distributional similarity is measured with Fréchet Audio Distance (FAD) [19] using two pretrained encoders: VGGish (FAD-VGGish) and CLAP [8][15] (FAD-CLAP). FAD-VGGish captures coarse timbre, instrument mix, and dynamics in a classification embedding space, whereas FAD-CLAP operates in an audio–text space emphasizing higher-level attributes such as style, instrumentation, and mood.

Metrics are computed on 5-second excerpts with 2-second overlap from original audio, audio generated by *Vocos* alone, *Vocos* with VAE, and the full motion-to-audio pipeline. All metrics are evaluated per dancer and training regime; ms-stft and ms-log-mel use only in-domain data, while FAD-VGGish and FAD-CLAP use both in-domain and out-of-domain data.

5.2 Qualitative Evaluation

Through critical listening and magnitude–spectogram inspection, this evaluation examined whether the audio generated by the different stages of the motion-to-audio pipeline (*Vocos*, *Vocos* & VAE, *Vocos* & VAE & Transformer) achieved sufficient perceptual quality For this evaluation, representative excerpts were selected from each recording: for Dancer1, different vocalisations in her father's voice (spoken words and non-verbal utterances); for Dancer2, stylistically distinct songs; and for Dancer3, acoustically contrasting sections within a single composition.

5.3 Artistic Evaluation

To assess the three pipeline variants in interactive, real-time use, each dancer participated in a second solo improvisation session using the models trained on their own movement and audio data. Dancers were free to choose movement vocabulary, exploratory strategies, and improvisation duration, and were not shown recordings of their initial sessions to encourage spontaneity. All sessions were documented with synchronised video and motion capture.

After each improvisation, dancers recorded spontaneous voice-memo reflections. Once all three sessions were completed, semi-structured interviews provided deeper insight into experiences, strategies, and perceptions, focusing on: (1) model preference (most/least engaging versions and reasons), (2) overall experience (relative to fixed music), (3) control and agency (felt influence of movement on sound), (4) sense-making and intuition (how dancers understood the mapping), (5) expressiveness and flow (whether sound reflected expressive movement and supported sustained absorption), and (6) artistic potential (possible integration into artistic practice and performance). The three variants referred to in the discussion are: Version 1 for training regime (i), Version 2 for training regime (ii), and Version 3 for training regime (iii).

6 Results

The results of the evaluations are presented in this section, complemented by online video and audio material.

6.1 Quantitative Evaluation

Across all datasets and training regimes, the VAE matches the *Vocos* baseline in reconstruction fidelity, with similar STFT and mel-spectral distances, while the full pipeline yields higher errors (see figure 4). Distributional results are more varied (see figure 5): the VAE often achieves lower FAD than *Vocos*, especially in VGGish space, indicating better alignment with global audio statistics. The transformer shows intermediate performance, sometimes outperforming *Vocos* but generally with higher distances and worse CLAP-based FAD, suggesting reduced semantic consistency.

Stronger regularisation slightly degrades reconstruction for both VAE and transformer, though modestly. Joint-dropout further worsens reconstruction but consistently improves distributional match for the transformer, particularly out–of–domain, where FAD in both VGGish and CLAP spaces is often lowest.

For Dancer1 and Dancer3, out–of–domain transformer models with joint-dropout show notably lower FAD than non–joint-dropout variants, while for Dancer2 the gains are weaker or may reverse.

			Dancer 1			Dancer2			Dancer3		
Data Domain	Audio Source	Metrics Type	β 0.1	β 0.1 & jd	β 1.0	β 0.1	β 0.1 & jd	β 1.0	β 0.1	β 0.1 & jd	β 1.0
In-Domain	vocos	ms-stft	1.143	1.143	1.143	1.165	1.165	1.165	1.279	1.279	1.279
In-Domain	vae	ms-stft	1.119	1.116	1.173	1.210	1.209	1.262	1.294	1.297	1.315
In-Domain	transformer	ms-stft	1.412	1.520	1.329	1.387	1.561	1.398	1.411	1.519	1.434
In-Domain	vocos	ms-log-mel	0.984	0.984	0.984	0.908	0.908	0.908	1.075	1.075	1.075
In-Domain	vae	ms-log-mel	0.922	0.918	0.976	0.927	0.926	1.000	1.062	1.066	1.094
In-Domain	transformer	ms-log-mel	1.274	1.408	1.176	1.130	1.345	1.155	1.208	1.318	1.239

Figure 4: Quantitative analysis of reconstruction fidelity. The acronym *jd* stands for joint-dropout.

			Dancer1			Dancer2			Dancer3		
Data Domain	Audio Source	Metrics Type	β 0.1	β 0.1 & jd	β 1.0	β 0.1	β 0.1 & jd	β 1.0	β 0.1	β & jd	β 0.1
In-Domain	vocos	fad-vggish	1.866	1.866	1.866	2.185	2.185	2.185	2.723	2.723	2.723
In-Domain	vae	fad-vggish	0.515	0.528	1.013	1.713	1.703	2.465	0.935	0.982	1.414
In-Domain	transformer	fad-vggish	0.910	1.262	1.356	2.044	6.496	3.941	1.485	2.390	1.764
Out-of-Domain	transformer	fad-vggish	3.485	3.108	4.488	9.559	7.900	10.773	4.195	2.406	4.166
In-Domain	vocos	fad-clap	0.179	0.179	0.179	0.119	0.119	0.119	0.172	0.172	0.172
In-Domain	vae	fad-clap	0.108	0.100	0.252	0.151	0.153	0.277	0.181	0.185	0.295
In-Domain	transformer	fad-clap	0.157	0.317	0.233	0.245	0.640	0.430	0.327	0.464	0.342
Out-of-Domain	transformer	fad-clap	0.825	0.592	0.972	0.960	0.836	0.973	0.608	0.487	0.675

Figure 5: Quantitative analysis of distributional match. The acronym *jd* stands for joint-dropout.

6.2 Qualitative Evaluation

Reconstruction characteristics varied in background noise, perceptual quality, and high-frequency detail. For Dancer1, *Vocos* slightly reduces background noise[10] present in the original[11]. For Dancer2, original[12] and *Vocos* resynthesis[13] are perceptually very similar, whereas for Dancer3, *Vocos* departs from the glitchy broadband-noise original[14], imparting a fluttering, modulated-noise character[15].

Reconstruction also depends on β. For Dancer1, both $\beta = 0.1$ and $\beta = 1.0$ produce intelligible speech with increased granularity, tonal noise, and some loss of high frequencies, more pronounced at high β[16] [17]. For Dancer2, both variants retain high perceptual fidelity, preserving timbre and temporal structure with mild spectral fluctuations and slightly smoothed high frequencies[18] [19]. For Dancer3, the *Vocos*-induced fluttering noise persists for both β values[20] [21], alongside attenuation and loss of high-frequency components.

[10] Audio: Dancer1, *Vocos*
[11] Audio: Dancer1, Original
[12] Audio: Dancer2, Original
[13] Audio: Dancer2, *Vocos*
[14] Audio: Dancer3, Original
[15] Audio: Dancer3, *Vocos*
[16] Audio: Dancer1, VAE β=0.1
[17] Audio: Dancer1, VAE β=1.0 Decoder-only Post-training
[18] Audio: Dancer2, VAE β=0.1
[19] Audio: Dancer2, VAE β=1.0 Decoder-only Post-training
[20] Audio: Dancer3, VAE β=0.1
[21] Audio: Dancer3, VAE β=1.0 Decoder-only Post-training

The motion-to-audio pipeline was evaluated for in–domain and out–of–domain motion-to-audio generation along perceptual quality, temporal alignment, and the effect of joint-dropout. In–domain, Dancer1 and Dancer2 models exceed VAE-only quality, with the Transformer compensating VAE artefacts (reducing tonal noise and high-frequency loss for Dancer1, preserving musical fidelity for Dancer2). For Dancer3, in–domain generation without joint-dropout slightly improves on the VAE alone but still lacks the original broadband noise, instead retaining the *Vocos* fluttering texture[22] [23]. Out–of–domain, audio quality degrades for Dancer1 and Dancer2, with increased granularity and noise (especially without joint-dropout), while for Dancer3 it remains close to in–domain, likely because additional artefacts are masked by the glitch-like source material[24].

Temporal alignment follows a similar pattern. In–domain, temporal coherence is generally high without joint-dropout[25] [26] [27]. With joint-dropout, small temporal shifts appear for Dancer1 and more pronounced timing degradation for Dancer2, especially in strongly pulsed styles[28] [29]. For Dancer3, in–domain alignment remains high even with joint-dropout, though temporal smearing of the noisy texture becomes audible[30]. Out–of–domain, temporal coherence deteriorates strongly for Dancer1 without joint-dropout (prosody and timbre are retained but lexical intelligibility is largely lost), whereas joint-dropout maintains better temporal consistency and often intelligible speech[31] [32]. For Dancer2, all models show reduced temporal precision and elevated noise out–of–domain, but musical styles remain recognizable, appearing as shorter style-switching intervals without joint-dropout and longer, more stable segments with it[33] [34].

Overall, these findings highlight the role of joint-dropout in balancing cross–modal coupling and intra–audio structure. For Dancer1 and Dancer2, it reduces fine-grained in–domain alignment but improves temporal organization and robustness out–of–domain, yielding more intelligible speech for Dancer1 and more temporally and stylistically consistent segments for Dancer2. For Dancer3, the same mechanism can yield aberrant out–of–domain generations with sustained tonal bands and reduced resemblance to the glitch-style source, suggesting that for highly stochastic audio the Transformer may struggle to learn meaningful audio-only structure under joint-dropout and may decouple from motion-driven behaviour[35].

6.3 Artistic Evaluation

Short online video excerpts are provided for all improvisation sessions.[36] [37] [38] [39] [40] [41] [42] [43] [44]

The following results synthesize the dancers' feedback to the interview questions.

Model preference: The dancers articulated distinct preferences: Dancer1 favoured Version 2 for its recognizable speech fragments and strong affective connection to her father's voice; Dancer3 was most drawn to Version 3, stressing a sense of "play" and co-agency in exploring audio variations; Dancer2 clearly preferred Version 1, where she perceived a direct relation between movement and sound and could "feel" her body producing sound.

Overall experience: All dancers described the sessions as engaging but cognitively and physically demanding. Dancer1 framed her work as an ongoing search for rules that never fully stabilized, oscillating between rewarding moments and frustration. Dancer3 experienced the shift from complementing a fixed score to co-creating sound as deeply rewarding yet effortful due to continuous exploration. Dancer2 particularly enjoyed moments when the movement–sound connection felt clear, enabling her to revisit existing performance material.

Control and agency: Perceptions of control and agency were mixed and dependent on the model variant. Across all interviews, dancers reported their strongest sense of agency when a movement–music relationship became repeatable and thus learnable: for Eleni in Version 1, by re-using remembered movement and music relationships from the original performance; for Diane in Version 2, by discovering movements that repeatedly brought her father's voice back into intelligible fragments; and for Tim in Version 3, by finding specific poses or rhythmic movement patterns that reliably produced similar sonic outcomes. Conversely, a low sense of agency emerged whenever such repeatability could not be established, or when dancers searched for but failed to discover clear causal relationships between movement and music, such as the expectation that highly energetic movement should yield highly energetic music (Tim).

Sense-making and intuition: All dancers used iterative, exploratory tactics: systematically varying geometry, speed, and spatial relations; alternating analytic probing with open improvisation; and progressively reducing the number of movement parameters to isolate effective movement effects. Over time, Dancer2 developed an intuitive grasp of Version 1, whereas Dancer1 and Dancer3 reported transient moments of understanding that repeatedly dissolved, keeping global predictability low.

Expressiveness and flow: Experiences of expressiveness and flow depended on mapping legibility. Dancer1 and Dancer3 noted that conventional expectations (e.g., "more expressive" movement yielding more intense sound) were not reliably met, limiting their ability to sculpt musical expressivity and yielding only brief, fragile flow states. Dancer2, in contrast, described Version 1 as supporting a relatively continuous flow once she had established an embodied understanding of how movement qualities affected sound.

[22] Motion-Audio: Dancer3, Original Audio
[23] Motion-Audio: Dancer3, In–Domain, β=0.1
[24] Motion-Audio: Dancer3, Out–of–Domain, β=0.1
[25] Motion-Audio: Dancer1, In–Domain, β=0.1
[26] Motion-Audio: Dancer2, In–Domain, β=1.0
[27] Motion-Audio: Dancer3, In–Domain, β=1.0
[28] Motion-Audio: Dancer1, In–Domain, β=0.1, joint-dropout
[29] Motion-Audio: Dancer2, In–Domain, β=0.1, joint-dropout
[30] Motion-Audio: Dancer3, In–Domain, β=0.1, joint-dropout
[31] Motion-Audio: Dancer1, Out–of–Domain, β=0.1
[32] Motion-Audio: Dancer1, Out–of–Domain, β=0.1, joint–dropout
[33] Motion-Audio: Dancer2, Out–of–Domain, β=0.1
[34] Motion-Audio: Dancer2, Out–of–Domain, β=0.1, joint-dropout
[35] Motion-Audio: Dancer3, Out–of–Domain, β=0.1, joint-dropout

[36] Video Dancer1, Version1
[37] Video Dancer1, Version2
[38] Video Dancer1, Version3
[39] Video Dancer2, Version1
[40] Video Dancer2, Version2
[41] Video Dancer2, Version3
[42] Video Dancer3, Version1
[43] Video Dancer3, Version2
[44] Video Dancer3, Version3

Artistic potential: All dancers recognised substantial artistic potential, albeit in different directions. Dancer1 emphasized a dramaturgy of searching, loss of control, and the precarious re-animation of her father's voice. Dancer3 foregrounded the system's capacity to sonify micro--parameters of movement, with implications for performance, pedagogy, and possibly somatic or rehabilitative contexts. Dancer2 framed the system as a co-creator capable of generating soundscapes directly from choreography, aligning with her interest in rebalancing authorship and co-agency between sound and movement.

7 Discussion

The development of new model variants for translating a dancer's movement into music and the evaluation of these variants has led to important insights that will continue to guide the technical and artistic aspects of the work. In the following, the results obtained from the quantitative and qualitative assessment of models' performance as well as the dancers feedback provided after their real-time engagement with the models are discussed.

7.1 Quantitative Evaluation

The VAE preserves most spectral structure, adding only small reconstruction error relative to the *Vocos* baseline. The Transformer generates reasonably faithful audio tokens from motion, and its higher reconstruction errors mainly reflect the greater difficulty of motion-guided generation versus direct reconstruction. The VAE also regularizes *Vocos*, indicating that choosing *Vocos* trades off processing speed against audio quality, while Transformer-induced errors remain below the *Vocos* margin and thus do not dominate overall degradation. Consequently, *Vocos* contributes a substantial portion of the mismatch between Transformer outputs and real audio, suggesting that improving or removing the vocoder may yield larger gains than further optimizing the VAE or Transformer.

Training-regime variations reveal a consistent trade-off between reconstruction fidelity and distributional alignment. The custom joint-drop regime exemplifies this: it typically causes modest degradation in Transformer reconstruction metrics while systematically improving out–of–domain FAD in both VGGish and CLAP spaces. This underscores the need for revised training regimes that better balance reconstruction quality and distribution match.

Dancer–specific results show that these trade–offs are not uniform across datasets. For Dancer1 and Dancer2, regimes with the best out–of–domain distributional metrics also match the dancers' preferred *PEMIDA* versions in subjective evaluations, suggesting that improved distributional alignment is reflected in perceived quality.

7.2 Qualitative Evaluation

The qualitative evaluation shows that the VAE learns useful latent representations and that the Transformer successfully models cross-modal audio–motion dependencies even from very small, dancer-specific datasets. This makes the architectures suitable for exploratory artistic workflows with dancer-specific material and rapid iteration, but not yet adequate for professional production, given limited audio fidelity—especially for out-of-domain motion-to-audio—and occasional aberrant Transformer behaviour when musical coherence is favoured over motion–audio correspondence. The most salient findings concern the effect of *Vocos* and differences between in--domain and out--of--domain Transformer behaviour, particularly with joint–dropout.

Vocos quality varies markedly: for spoken voice (Dancer1) and popular-music idioms (Dancer2), mel-spectrogram resynthesis is almost indistinguishable from the originals, whereas for noisy glitch material (Dancer3) it introduces pronounced fluttering artefacts that substantially degrade quality. Because the 48 kHz *Vocos* training data are undocumented, it remains unclear whether this reflects a training bias or an inherent limitation of the inverse short-time Fourier transform configuration.

Transformer models without joint-dropout perform well in both motion–audio alignment and overall audio quality, sometimes compensating for VAE decoder artefacts, indicating that motion-to-audio mappings can be learned from limited dancer-specific data. Out-of-domain evaluation, however, reveals pronounced dataset differences: speech intelligibility collapses without joint-dropout for Dancer1, artefacts increase substantially for Dancer2, while the glitch aesthetics of Dancer3 remain largely unaffected. This limited generalisation raises questions about how "generalisation" should be understood in performer-specific, artist-driven systems and under which artistic objectives strong generalisation is desirable or detrimental.

The custom joint-dropout configuration introduces further dataset-dependent effects. Designed to increase robustness to motion deviations and reduce the number of joints required for control, it inadvertently increases temporal consistency in the generated audio at the cost of precise motion–audio alignment. This trade-off is beneficial when temporal coherence in audio is crucial (e.g., spoken text in Dancer1) but detrimental for audio that relies on strict periodic structure (e.g., metrically regular popular music in Dancer2).

7.3 Artistic Evaluation

Involving dancers in evaluating the motion-to-audio pipeline through free improvisation and post-hoc interviews was essential for understanding how the system supports creative practice and aligns with individual artistic interests.

Four main themes emerged: diverging preferences for model variants; tension between dancer agency and the opacity of the motion–audio mapping; shared strategies of sense-making and difficulties sustaining flow; and distinct yet converging visions of the system's artistic potential. Each dancer preferred a different variant, reflecting specific aesthetic and dramaturgical priorities (e.g., recognisable voice fragments and affective resonance for Dancer1, nuanced audio variation via micro--movement for Dancer3, clearly legible movement–music relations for Dancer2), indicating there is no universally optimal model behaviour but rather dependence on material and artistic approach.

All dancers perceived a relationship between movement and music but struggled to infer stable, global mapping principles. Perceived control and agency were strongly model- and dancer--dependent, yet a consistent pattern emerged: agency increased when movement–music relationships were sufficiently repeatable to support robust expectations. At the same time, much of their frustration and sense of algorithmic opacity seemed to arise from a mismatch between dancers' implicit control metaphors and the structures actually learned from the training data, which did not necessarily follow these rules. This suggests that dancers and choreographers need greater awareness of the implicit rules they themselves enact when working with movement and music

during data collection, so that expectations, training material, and model behaviour can be better aligned.

During improvisation, dancers shifted from unstructured trial and error to more systematic exploration, testing hypotheses, probing the mapping, isolating and recombining movement parameters, and alternating analytic and intuitive modes. This process helped them uncover local movement–sound relationships and develop situated strategies that enabled access to flow states. Flow typically coincided with local periods of apparent alignment between movement and sound; two dancers emphasised the fragility of these states, whereas the third reported a more continuous sense of reciprocal expressiveness once she had internalised the system's behaviour. At the same time, they valued the system's partial autonomy and the creative potential of its non-determinism, highlighting a delicate balance between controllability and independence. Overall, the observations point to a persistent tension between felt agency and algorithmic opacity, indicating that clearer movement–sound relationships are desirable but must be carefully balanced against preserving system autonomy.

All dancers expressed a wish to continue working with the system and to integrate it into future projects, outlining dancer--specific trajectories: Dancer1 focusing on autobiographical material and her relationship to her father's recorded voice; Dancer3 using motion-controlled audio as auditory feedback for micro--movement in training and pedagogy; and Dancer2 extending choreography so that composing movement simultaneously becomes music making. At the same time, they converged on positioning the system as a co-creative partner rather than a mere technical tool, foregrounding engagement with non-human agency and AI-mediated serendipity in dance and music creation. Taken together, these insights indicate that all three dancers recognise substantial artistic potential in *PEMIDA*, and that the system is flexible enough to accommodate divergent artistic strategies.

8 Outlook

The current *PEMIDA* system has proven technically viable and has sparked considerable artistic interest among the participating dancers, yet there remains substantial scope for refinement in model architectures, training strategy, and integration into artistic practice.

8.1 Technical Revisions and Improvements

Using *Vocos* for preprocessing reduces dimensionality and makes audio more amenable to representation learning and motion–audio correlation modelling, but introduces noticeable artefacts for noisy, textural, and glitch-oriented music. One option is to partially retrain *Vocos* on the dancers' own music via transfer learning, although the limited material may cap improvements; alternatively, it could be replaced or complemented by full-band 48 kHz neural audio codecs such as *EnCodec* [7] and *AudioDec* [37], or audio autoencoders like *Music2Latent* [27] designed for downstream tasks.

The current simple VAE trains quickly on small datasets but falls short of state-of-the-art reconstruction quality compared to specialised models such as *RAVE*. Planned improvements therefore include revising the architecture and training regime (e.g., adopting RAVE's separated/downsampled high-frequency subbands and a decoder-only adversarial fine-tuning stage) and experimenting with accelerated *RAVE* training by first using the combined corpus and then refining on each dancer's material.

In the current design, the Transformer must learn both temporal alignment and instantaneous feature mappings between motion and audio; the latter could be eased by an explicit cross-modal feature learning stage, where self-supervised models first learn modality-specific representations and align motion and audio in a shared embedding space so that corresponding segments lie close and motion features directly guide audio generation. Within *PEMIDA*, the Transformer could then exploit this semantic proximity for more reliable, nuanced motion-to-audio translation, building on existing work on cross-modal feature learning for motion and audio (e.g. [23], [25], [38]).

8.2 Artistic Applications

Evaluations with dancers reveal a need to enhance *PEMIDA*'s legibility and learnability while preserving a productive balance between dancer agency and system autonomy, since clear mappings are crucial for sustaining flow states. Mapping legibility could be increased through more constrained and interpretable mappings (e.g., specifying which joints or body regions influence sound, or restricting which latent regions a given movement configuration can access) and by adapting improvisation strategies during data collection to favour reduced diversity and systematic repetition so models can detect robust motion–music patterns.

The findings also underscore the value of multiple model variants, as dancers responded differently to specific configurations, implying that architectures, training settings, and interaction setups should be tailored to each dancer's aesthetic and dramaturgical priorities. This points toward highly personalised systems customised not only via dancer-specific training data but also through differentiated model designs and interaction modes. A crucial next step is to reduce hardware requirements so dancers can work independently, replacing the Xsens system with accessible alternatives (e.g., consumer cameras plus 3D pose estimation or simple wireless sensors for selected body parts) and simplifying *PEMIDA*'s computational demands so it can run reliably on standard, non-GPU machines.

Finally, both ML researchers and dancers intend to integrate *PEMIDA* into new choreographic and performance projects and to collaborate on funding applications to support further development and production.

9 Acknowledgments

The authors are grateful to the Immersive Arts Space at the Zurich University of the Arts (ZHdK) for generously lending their Xsens motion capture system for extended periods.

10 Ethical Standards

This work was funded exclusively through internal support from ZHdK, and all collaborating dancer co-authors provided informed consent for their participation.

References

[1] Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. 2015. Scheduled sampling for sequence prediction with recurrent neural networks. *Advances in neural information processing systems* 28 (2015).

[2] Andreas Bergsland. 2022. Designing interactive sonifications for the exploration of dance phrase edges. In *Proceedings of the 19th Sound and Music Computing Conference*. SMC Network.

[3] Frédéric Bevilacqua, Lisa Naugle, and Isabel Valverde. 2001. Virtual dance and music environment using motion capture. In *Proc. of the IEEE-Multimedia Technology And Applications Conference, Irvine CA*, Vol. 2.

[4] Daniel Bisig and Kıvanç Tatar. 2021. Raw music from free movements: Early experiments in using machine learning to create raw audio from dance movements. In *Proc. 2nd Conf. AI Music Creativity*.

[5] Antoine Caillon and Philippe Esling. 2021. RAVE: A variational autoencoder for fast and high-quality neural audio synthesis. *arXiv preprint arXiv:2111.05011* (2021).

[6] Luke Dahl, Christopher Knowlton, and Antonia Zaferiou. 2019. Developing real-time sonification with optical motion capture to convey balance-related metrics to dancers. In *Proceedings of the 6th International Conference on Movement and Computing*. 1–6.

[7] Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. 2022. High fidelity neural audio compression. *arXiv preprint arXiv:2210.13438* (2022).

[8] Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. 2023. Clap learning audio concepts from natural language supervision. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 1–5.

[9] Jesse Engel, Lamtharn Hantrakul, Chenjie Gu, and Adam Roberts. 2020. DDSP: Differentiable digital signal processing. *arXiv preprint arXiv:2001.04643* (2020).

[10] Jerry Alan Fails and Dan R Olsen Jr. 2003. Interactive machine learning. In *Proceedings of the 8th international conference on Intelligent user interfaces*. 39–45.

[11] Frey Faust. 2011. *The Axis Syllabus: Human Movement Lexicon*. Axis Syllabus Research Community, n.p. Living, continuously updated edition.

[12] Jules Françoise and Frederic Bevilacqua. 2018. Motion-sound mapping through interaction: An approach to user-centered design of auditory feedback using machine learning. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 8, 2 (2018), 1–30.

[13] Emma Frid, Ludvig Elblaus, and Roberto Bresin. 2019. Interactive sonification of a fluid dance movement: an exploratory study. *Journal on Multimodal User Interfaces* 13, 3 (2019), 181–189.

[14] Hao Fu, Chunyuan Li, Xiaodong Liu, Jianfeng Gao, Asli Celikyilmaz, and Lawrence Carin. 2019. Cyclical annealing schedule: A simple approach to mitigating kl vanishing. *arXiv preprint arXiv:1903.10145* (2019).

[15] Azalea Gui, Hannes Gamper, Sebastian Braun, and Dimitra Emmanouilidou. 2024. Adapting frechet audio distance for generative music evaluation. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 1331–1335.

[16] Lamtharn Hantrakul. 2018. GestureRNN: A neural gesture system for the Roli Lightpad Block. In *Proceedings of the International Conference on New Interfaces for Musical Expression*. 132–137.

[17] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. 2017. beta-vae: Learning basic visual concepts with a constrained variational framework. In *International conference on learning representations*.

[18] Reynald Hoskinson, Kees van den Doel, and Sidney S Fels. 2003. Real-time Adaptive Control of Modal Synthesis.. In *NIME*. 99–103.

[19] Kevin Kilgour, Mauricio Zuluaga, Dominik Roblek, and Matthew Sharifi. 2018. Fr\'echet audio distance: A metric for evaluating music enhancement algorithms. *arXiv preprint arXiv:1812.08466* (2018).

[20] Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* (2013).

[21] Steven Landry and Myounghoon Jeon. 2017. Participatory design research methodologies: A case study in dancer sonification. (2017).

[22] James Leonard and Andrea Giomi. 2020. Towards an interactive model-based sonification of hand gesture for dance performance. In *Proceedings of the International Conference on New Interfaces for Musical Expression*. 369–374.

[23] Fan Liu, De-Long Chen, Rui-Zhi Zhou, Sai Yang, and Feng Xu. 2022. Self-supervised music motion synchronization learning for music-driven conducting motion generation. *Journal of Computer Science and Technology* 37, 3 (2022), 539–558.

[24] Julien Mercier and Irini Kalaitzidi. 2024. Motion capture data as machine-readable notation to capture musical interpretation: experimenting with movement sonification and synthesis. In *Proceedings of the Ninth International Conference on Technologies for Music Notation and Representation (TENOR)*, Vol. 2024. Zurich University of the Arts, 169–171.

[25] Joseph Meyer, Nick Bryan-Kinns, Sarah Fdili Alaoui, Mick Grierson, and Rebecca Fiebrink. 2025. Interactive Movement-to-Audio with Pre-Trained Neural Networks. In *Proceedings of the 2025 Conference on Creativity and Cognition*. 491–493.

[26] Sarah Nabi, Philippe Esling, Geoffroy Peeters, and Frédéric Bevilacqua. 2024. Embodied exploration of deep latent spaces in interactive dance-music performance. In *Proceedings of the 9th International Conference on Movement and Computing*. 1–9.

[27] Marco Pasini, Stefan Lattner, and George Fazekas. 2024. Music2latent: Consistency autoencoders for latent audio compression. *arXiv preprint arXiv:2408.06500* (2024).

[28] David Rokeby. 1998. The construction of experience: Interface as content. *Digital Illusion: Entertaining the future with high technology* 27 (1998), 47.

[29] Hubert Siuzdak. 2023. Vocos: Closing the gap between time-domain and fourier-based neural vocoders for high-quality audio synthesis. *arXiv preprint arXiv:2306.00814* (2023).

[30] Christian J Steinmetz and Joshua D Reiss. 2020. auraloss: Audio focused loss functions in PyTorch. 124.

[31] Vanessa Tan, Junghyun Nam, Juhan Nam, and Junyong Noh. 2023. Motion to dance music generation using latent diffusion model. In *SIGGRAPH Asia 2023 Technical Communications*. 1–4.

[32] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).

[33] Olga Vechtomova and Jeff Bos. 2025. Reimagining Dance: Real-time Music Co-creation between Dancers and AI. *arXiv preprint arXiv:2506.12008* (2025).

[34] Gabriel Vigliensoni, Rebecca Fiebrink, et al. 2023. Steering latent audio models through interactive machine learning. (2023).

[35] Todd Winkler. 1997. Creating interactive dance with the very nervous system. In *Proceedings of Connecticut College Symposium on Arts and Technology*, Vol. 2.

[36] Matthew Wright, Adrian Freed, et al. 1997. Open soundcontrol: A new protocol for communicating with sound synthesizers. In *ICMC*.

[37] Yi-Chiao Wu, Israel D Gebru, Dejan Marković, and Alexander Richard. 2023. Audiodec: An open-source streaming high-fidelity neural audio codec. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 1–5.

[38] Jiashuo Yu, Junfu Pu, Ying Cheng, Rui Feng, and Ying Shan. 2023. Learning music-dance representations through explicit-implicit rhythm synchronization. *IEEE Transactions on Multimedia* 26 (2023), 8454–8463.

[39] Qiushi Zhou, Cheng Cheng Chua, Jarrod Knibbe, Jorge Goncalves, and Eduardo Velloso. 2021. Dance and choreography in HCI: a two-decade retrospective. In *Proceedings of the 2021 CHI conference on human factors in computing systems*. 1–14.