Con Moto: Embodied Steering of Music Transformers for Live Dance Improvisation

Zhixing Chen*
zhixingc@mit.edu
Massachusetts Institute of Technology
Cambridge, USA

Heidi Lei*
hlei@mit.edu
Massachusetts Institute of Technology
Cambridge, USA

Cheng-Zhi Anna Huang
huangcza@mit.edu
Massachusetts Institute of Technology
Cambridge, USA

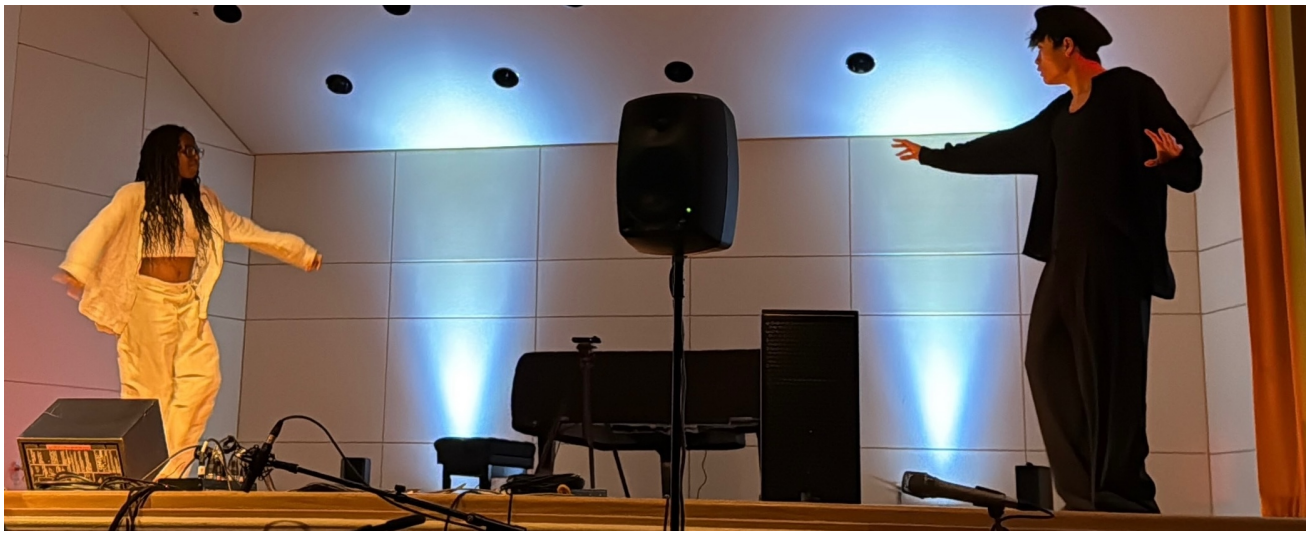

Figure 1: Dancers performing with real-time music accompaniment generated by *Con Moto*.

Abstract

Con Moto is a real-time generative music system for dance improvisation that supports embodied steering of a transformer model with configurable levels of agency. While existing frameworks demonstrate the potential of embodied music-making and movement sonification in live performance, achieving both high musical coherence and low-latency responsiveness remains an ongoing challenge. In response, we leverage the musical coherence of real-time MIDI-based transformer models to design an integrated system that translates camera motion data into movement parameters, which in turn control the musical output. *Con Moto* employs two layers of control strategies: 1) inference-time steering of the transformer model and 2) post-generation rendering using Max/MSP as a control interface and Ableton Live for sound synthesis. We present the system through a duet performance for a live audience, supplemented by qualitative reflections from the dancers and the audience. By reconfiguring how the dancers' movements map to musical functions, we create a system with configurable agency. Fine-grained control over an individual musical voice invited dancers to experience the system as a playable instrument, while abstract musical mappings to a genre's energy opened space for the system to act as an autonomous creative partner. By navigating the aesthetic friction between human intent and AI agency, we explore a dynamic that facilitates a deep, bidirectional feedback loop.

Keywords

Embodied Music interaction, Human-AI Co-creation, Controllable music generation, Real-time controls, Movement-based interaction, Inference-time Model Steering

1 Introduction

Live dance performance is a deeply embodied experience, in which dancers interpret and express musical structure through physical movement. Timing, weight, spatial trajectories, and gesture all reflect a dancer's understanding of the music as it unfolds, and these elements are shaped by both anticipation and reaction. In this sense, dance is not merely synchronized to sound, but is an active, perceptual engagement with musical form, phrasing, and energy.

Within improvisational traditions—often referred to as "freestyling" within Hip-Hop and street dance cultures— dancers continuously listen to the music, form expectations about its future evolution, and adapt their movements in response to deviations, accents, and emergent structure. Dancers rely on their musicality to navigate the feedback loop between perception and action. In this work, we use the terms "improvisation" and "freestyling" interchangeably to honor the diverse contemporary and street-style lineages that inform our practice.

We are interested in extending the dance-music reciprocal dynamic to human–AI collaboration [8], exploring settings in which dancers do not only respond to music, but also participate

*Both authors contributed equally to this research.

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

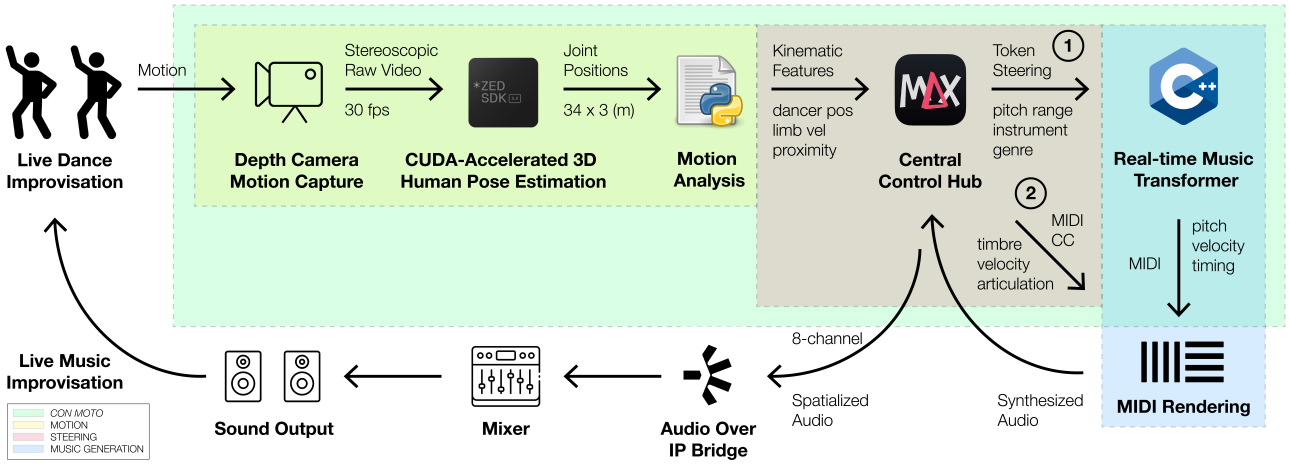

Figure 2: The system is designed to facilitate collaboration between human dancers and a generative AI in a live setting. The dancers' physical motion is captured through a camera and processed into control signals, consumed both by the generative model outputting plain MIDI events, and the DAW rendering the MIDI. The synthesized audio is played live to the dancers, who simultaneously react to the generated music and shape the future generation through motion.

in shaping its generation in real time. Rather than treating music as a fixed or externally determined stimulus, we ask how generative systems might support a shared improvisational space in which movement and sound co-evolve. Such a framing positions the generative model not as an accompaniment engine, but as a responsive, high-agency partner in performance [5, 6, 10]. Our approach follows a first-person Dialogic Design process [29], moving away from hierarchical designer-user roles toward a multi-directional flow of ideas between our dancer-researcher and AI-researcher authors. Over two months of weekly design and movement sessions, mangling artistic and technical constraints [23], we expanded this collaboration to include an external professional dancer to investigate how AI can mediate human-human resonance in addition to the individual human-AI loop.

Designing for this kind of collaboration introduces a central tension between control and autonomy. If dancers are given direct, low-level control over musical parameters, the system becomes a conventional instrument with limited generative agency. Conversely, if the model operates with too much autonomy, the dancer's influence becomes opaque, weakening the sense of authorship and intentionality. Achieving a productive balance requires interfaces and models that allow dancers to meaningfully influence musical behavior while preserving the system's capacity for independent musical decision-making.

We introduce *Con Moto*, a system that explores this balance by coupling embodied movement input with real-time generative music processes, aiming to support collaborative improvisation between dancers and an AI-driven musical partner. To evaluate the system's expressive potential, we present *unscored* (Figure 1), a live improvisation performance in which two dancers navigate a shared sonic space to steer both low-level nuances—such as timbre, velocity, and articulation—and high-level structural parameters, including harmony, instrumentation, and genre[1].

The system consists of a real-time generative music model coupled with a motion-based control pipeline designed for live performance (Figure 2). A motion-capture camera tracks the dancer's location and movement on stage, extracting spatial and activity-related features from the performer's motion. These features are mapped to musical control signals that influence the behavior of the generative model and downstream sound synthesis processes. In particular, spatial position is mapped to high-level pitch constraints, such as the active pitch range, while measures of movement intensity and the detection of specific movement patterns are mapped to timbral parameters via a digital audio workstation and a Max/MSP-based control layer. The generative model operates continuously. It regenerates musical material in response to incoming control messages including changes in pitch range or instrument availability, allowing the dancer's movement to shape both the harmonic space and the sonic character of the music in real time.

Our key contributions are:

- An inference-time control strategy for steering a generative music model during live performance.
- An integrated system that enables movement-based control of music generation in a live setting, combining a motion-capture camera, a generative model, and mappings from embodied movement to control signals.
- Qualitative insights from both performers and audiences into a novel mode of music–dance interaction.

2 Related Work

We situate our work at the intersection of several domains: movement sonification, generative motion-to-music systems, real-time language model steering, and human-AI co-creation. Integrating high-level movement qualities, topological sonification methods, and diffusion-based generation directs the focus toward systems that do not merely respond to a performer, but resonate with them.

2.1 Movement sonification

Sonification is traditionally defined as the functional transformation of data and interactions into sound [13]. In dance, the transformation typically involves translating kinetic data—such as joint orientation, velocity, or acceleration—into musical parameters [2, 11, 15, 18]. Current approaches move beyond these structural mappings to treat sonification as a medium for what

[1] https://con-moto-nime.github.io/

Bevilacqua et al. [4] describes as sensori-motor enhancement. We ground our research on the notion that music is movement, viewing sound as not a secondary response to motion, but as an extension of the body's expressive intent [2].

Building on the work of Bang et al. [3], our approach utilizes a methodology of designing in proximity, prioritizing slowness and humility in the design process. This aligns with the framework of Technical Practice Research [22], where we treat our weekly studio sessions not as mere evaluations, but as the primary site of knowledge production. Rather than seeking a perfect technical mapping, we focus on the entanglement of sound and movement as it evolves through practice. In this state, the system facilitates a dual awareness: an inward-looking sensitivity to physical proprioception and an outward-looking poetic exploration of the sonic environment. Our goal is to leverage these affordances to move from reactive sonification toward a state of organic human-AI resonance.

2.2 Neural motion-to-music generation

Recent developments in neural audio synthesis have shifted the focus from deterministic mapping to generative systems. Early work by Aggarwal and Parikh [1] inverting the conventional paradigm—where motion is typically a response to audio—demonstrated the viability of using deep neural networks to synthesize music directly from dance. Researchers coin generative AI models that support interactions translating dance to music as "motion-to-music models."

Recent interactive performances have explored how navigating the latent spaces of generative models can facilitate expressive co-creation. Nabi et al. [21] utilizes the RAVE (Real-time Audio Variational Embedding) architecture, leveraging its capacity for low-latency effective timbre transfer [7]. By mapping Inertial Measurement Unit data directly to RAVE's latent space, this approach enables dancers and musicians to co-create in a shared sonic environment. Building on this, Meyer et al. [19] replaces wearable sensors with computer vision, training lightweight autoencoders to map video input directly to sound in real time. While these systems excel at reactive sound production, they often lack the structural complexity found in traditional musical compositions.

To address musical structure, diffusion-based methods have introduced tailored architectures for aligning motion and audio spaces, achieving higher musical quality and beat alignment scores than previous models [28, 30]. However, a critical limitation remains: these high-fidelity systems typically operate offline, as the iterative nature of diffusion processes struggles to meet the low-latency requirements of fluid live performance. Consequently, existing frameworks often face a trade-off between musical richness and real-time interactivity. Furthermore, while these works explore individual embodied music-making, the social and interpersonal dynamics between multiple dancers within these generative loops remain significantly underexplored.

2.3 Inference-time generation steering

The exploration is further grounded on the affordances of MIDI for real-time interaction, as demonstrated by the collaborative improvisations of Jordan Rudess and JAM_BOT [5]. While prior works on mapping motions or gestures to music often operate in the audio synthesis space enabled by real-time audio processing capabilities and thus often require precomposed music or expert-curated sound samples [20, 24, 26], generative models can offer more diversity and flexibility in the musical content.

Generating MIDI events instead of directly generating audio streams has multiple advantages. MIDI-based generation enables faster inference and significantly lower latency, while providing a more interpretable and controllable representation. Because musical structure is explicitly encoded at the event level, synthesis parameters can be tuned independently of the generative model and rendered through different instruments or timbral configurations. Moreover, inference-time constraints can be applied directly in the symbolic domain at substantially lower computational cost than in audio-domain models.

Prior work in natural language generation has explored sophisticated inference-time mechanisms for steering model outputs toward desired attributes, often relying on auxiliary classifiers, discriminators, or gradient-based updates during decoding [9, 16, 27]. These methods are designed to impose high-level semantic constraints in domains where structure is implicit and must be inferred from raw token sequences. In contrast, our approach operates in a symbolic musical domain where core structural attributes such as pitch, onset, duration are explicitly represented. This allows inference-time control to be implemented through simple, deterministic constraints on event selection, without auxiliary models or expensive optimization.

3 System design

The design of the *Con Moto* system is motivated by the pursuit of human-AI partnership, specifically investigating the continuum of danceability within generative music environments. While generative models offer a vast creative space, they often struggle with rhythmic consistency and structural predictability, appearing disconnected to a performer who requires a steady pulse to maintain an embodied dialogue. Consequently, our system design evolves around two key design considerations:

- Embedding high-level generative music models within intuitive, low-latency embodied control schemes.
- Applying generative heuristics to ensure the generated music remains rhythmically and structurally danceable.

The system is organized as a real-time control loop that couples embodied movement features to both symbolic music generation and audio rendering (Figure 2). *Con Moto* consists of three primary components: (1) a generative music model that produces symbolic musical material and supports inference-time steering via external control messages; (2) a depth camera and feature-extraction process that estimates the dancer's position, activity, and salient movement patterns from joint trajectories; and (3) a Max/MSP patch that serves as a central hub to receive OSC messages from Windows Laptop, send OSC messages to PC, send MIDI control signals to Ableton, receive multi-channel audio from Ableton, and send eight-channel spatialization to Dante Virtual Soundcard.

The generative model uses the anticipatory music transformer [25] as a pretrained backbone. The model is then finetuned on smaller datasets of specific music genres. One model is finetuned on the Maestro dataset[12], which contains expressive performance MIDI of classical piano pieces, to obtain a classical style model. We turn to a piano model to probe pitch range, leveraging the instrument's broad register as a space for expressive variation. Other models focus on dance-style music and are finetuned on

filtered subsets of the GigaMIDI dataset[17] including pieces labeled with genres such as jazz and disco. The real-time inference system for the generation is built upon the JAM_BOT system [5], which handles the model configuration, generation, and output scheduling in a multi-threaded system.

Motion data is streamed from the camera analysis process to Max/MSP via OSC. The analysis process extracts a compact set of features, including spatial location on stage, activity level, and detected movement motifs, which are transmitted at interactive rates. Within Max/MSP, these features are filtered, smoothed, and quantized to reduce jitter and to convert continuous movement into musically meaningful control signals.

Max/MSP formats and sends high-level control messages to the generative model, specifying constraints such as the active pitch range and the set of enabled instruments. The generative model runs in parallel and continuously produces symbolic musical output. When new control messages are received, the model regenerates or re-steers its output to reflect the updated constraints, allowing dancer movement to influence musical structure without requiring note-level control.

The symbolic output of the generative model is sent to the DAW for playback and MIDI-based synthesis. In parallel, Max/MSP maps movement-derived features to timbral and dynamic parameters in the DAW, enabling the dancer's movement to shape both the generated musical content and its sonic realization.

The system architecture is divided across three devices to balance GPU requirements and audio stability. Initial spatial tracking is performed on a Windows laptop (NVIDIA GPU). Tracking data is sent via OSC over a network to a MacBook running MaxMSP and Ableton Live. MaxMSP acts as the central hub, sending control signals to a high-performance PC running a native C++ music generation engine. Resulting MIDI data is returned to the MacBook for 8-channel spatialization and distributed to the mixer via Dante. This configuration addresses specific limitations in GPU compatibility and Windows audio driver reliability.

3.1 Pitch range and instrument control

To achieve inference-time control over pitch range and the active instrument, we apply a filtering operation to the model's note-selection process during generation (Figure 3). At each generation step, the model assigns scores to a set of possible next notes, i.e., a prior distribution $\pi(k)$, where $k \in \{0, \ldots, 127\}$ is a MIDI pitch number. Notes whose pitches or instrument classes fall outside of the currently active range $[p_{min}, p_{max}]$ are suppressed through masking, ensuring that only notes within the specified register can be selected. After masking we sample from the posterior distribution

$$\pi'(k) \propto \begin{cases} \pi(k) & \text{if } p_{\min} \le k \le p_{\max} \\ 0 & \text{otherwise} \end{cases}.$$

Since the instrument class is generated along with the pitch, the instrument control can be implemented by restricting the tokens to the desired instrument classes (Figure 4).

To reduce jitter and perceptual discontinuities caused by frequent regeneration, we apply several stabilizing strategies. First, we limit the rate of control updates so that a new pitch range or instrument configuration is only sent when there is a significant change in dancer input, defined in our performance as a shift of more than six semitones. Second, when regeneration is triggered, the system preserves the upcoming 250ms of already-generated

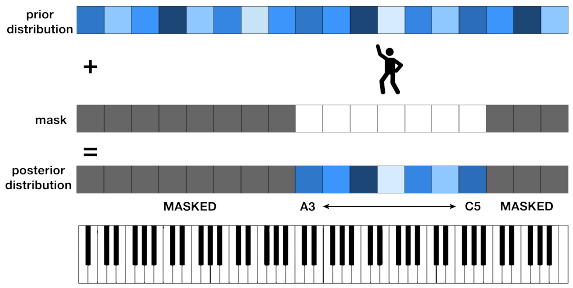

Figure 3: The pitch-token probabilities from the unconstrained model are masked to enforce a stage-position-dependent pitch range. Sampling is then performed from the masked posterior distribution.

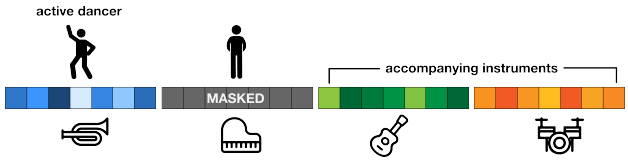

Figure 4: Similar to pitch range, instrument control is achieved through masking the non-active regions of the pitch-class token space.

events before discarding future output, allowing ongoing musical gestures to complete naturally. Finally, pitch constraints are updated smoothly rather than abruptly by interpolating between the previous and new pitch masks, enabling musical ideas initiated under the earlier constraint to resolve and resulting in more continuous, musically coherent transitions.

This design allows dancer movement to influence large-scale musical characteristics without requiring precise temporal alignment or discrete triggering. The generative model retains responsibility for choosing specific notes and rhythms within the constrained space, preserving stylistic consistency while remaining responsive to live, embodied input.

3.2 Harmonic control

Harmonic structure in the system is controlled through a learned representation of pitch-class activity rather than through explicit chord or scale constraints. We compute a chroma-based description of the musical context by aggregating the activation of the 12 pitch classes over sliding windows of one second. This representation captures which pitch classes are emphasized over time while remaining invariant to octave, providing a compact summary of harmonic content that is well aligned with human musical perception.

To integrate this information into generation, we extend the decoder transformer architecture with an additional encoder that processes the chroma representation. The encoded harmonic features are then injected into the generative model using cross-attention, allowing the model to condition its output on the current harmonic context. This design enables the model to adapt its note choices in response to evolving pitch-class distributions while preserving the model's ability to generate expressive and stylistically coherent musical material.

At performance time, we restrict the harmonic conditioning to a small set of interpretable states. Specifically, we select two chroma conditions corresponding to major and minor modal

profiles, which are extracted from real classical music rather than defined by explicitly activating a scale. These profiles are derived from pitch-class statistics computed over one-second windows of expressive classical performances, capturing characteristic harmonic tendencies and voice-leading patterns beyond simple scale membership.

3.3 Timbral, dynamic, and spatial control

In addition to steering the symbolic music generation in real time, *Con Moto* directly maps movement to the timbral, dynamic, and spatial layers of sound. Establishing a robust correspondence between movement and sound ensures that the performer's energy is reflected in the physicality of the music, thereby accentuating the dancer-music connection.

Movement velocity to MIDI velocity. The average limb velocity, calculated over a two-second sliding window, was mapped to MIDI note velocity. Velocity control established a foundational energy relation; if the dancers remained static, the generative engine produced no sound. This ensured that the musical output was a direct consequence of physical exertion, grounding the generation in embodied effort.

Verticality to articulation. The absolute height of the dancers were mapped to note duration. This mapping was co-designed with the dancers as a stable physical metaphor for controlling musical articulation—ranging from staccato movements at lower heights to sustained, legato tones when leaping or reaching upward.

Interpersonal proximity to distortion. To sonify the social dynamics between the performers, the spatial distance between the two dancers was mapped to vinyl distortion. As the dancers converged, the harmonic complexity and grit of the sound increased, providing a sonic representation of physical tension or intimacy.

Spatial Position to Sound Localization. The absolute XY coordinates of each dancer were mapped to virtual source positions within an 8-channel ambisonic field, implemented via the ICST Ambisonics plugin for Max/MSP. As dancers moved through the stage, the generative audio followed them, creating a localized sonic presence.

Our co-designed expressive control mappings are scaled across performance sections to support a dramatic arc. In the opening movements, velocity and position control a single instrument; by the finale, these parameters extend to map to EQ filters and the overall track volume, effectively opening the soundscape as the dancers' energy reaches its peak.

3.4 The co-design process

The development of *Con Moto* followed an iterative co-design methodology, driven by internal weekly sessions between the lead authors, supplemented by periodic insights from a professional dancer, musicians, artist-researchers, and sound designers; there were a few key turning points that shaped the evolution of the system. Across four sessions with the professional dancer, totaling five hours, we followed a soma design approach [14] to ensure the system's responses felt like a natural extension of the dancers' movements, treating their physical experience as the primary guide for refining the interaction.

Expanding embodied scale. While early prototypes focused on localized hand gestures over a piano, we realized that a larger, more expressive scale was needed to support full-body engagement and to explore an embodied experience of live steering of a generative model. Thus, we re-imagined the entire stage as a virtual piano, where mapping the dancer's spatial coordinates to musical range transformed the system from a localized tool into an expansive sonic space.

Defining danceability. The first session with the external dancer highlighted that pure generative agency can feel disconnected from a dancer's needs:

> *"... weird for me as a dancer, it's more like playing an instrument."*

To increase rhythmic legibility, we learned to force drum foundation at the start of generation and fine-tuned models on additional genres outside of classical (Jazz and Disco) with a steadier beat.

Configuring agency into narrative. To move beyond a feature showcase, we developed configurable agency tailored to a performance narrative. In our second session with the external dancer, we developed interpersonal mappings—such as proximity-based controls between dancers. In our session with other artist-researchers, they asked:

> *"Do you need to generate everything live?"*

While online generation offers a more interactive and adaptive experience given that the model generation can be steered in new directions through motion, we wanted to explore the capability of the system without the latency constraints by employing offline-generated tracks with fuller musical instrumentation in the last section of the performance.

Technical and timbral refinement. Sessions with musicians and sound designers prompted the addition of post-generation MIDI rendering and high-fidelity instrument packages, alongside spatialized audio to reinforce the physical connection between a dancer's location and the sound source.

4 Performance overview

To evaluate *Con Moto* in a multi-user context, the first-listed author—a dancer-researcher with a background in Hip-Hop and collegiate dance—partnered with an external dancer with background in ballet and Hip-Hop. We initially developed the system as a medium for the author's artistic self-expression, seeking to create a tool that resonated with their personal movement style and musical sensibilities. However, we also aimed to explore its potential as a collaborative platform. The two dancers shared a pre-existing creative rapport from their time on the same competitive dance team, which provided a foundational movement language and shared shorthand that significantly streamlined the co-design process.

The team collaboratively developed *unscored*, an improvised performance where the musical structure is entirely emergent. While the first-listed author served as a system architect and performer, the external dancer acted as an additional expert consultant, providing real-time and retrospective insights. Through the collaborative process, we refined the system to manage both low-level musical nuances—pitch, velocity, timbre, articulation, and tempo—as well as high-level concepts of harmony, instrumentation, and genre. The resulting performance for an audience of ~70 at Harvard University's Paine Concert Hall[2] is structured into two acts, each with two scenes, outlined below:

[2]https://music.fas.harvard.edu/event/huseac-farewell-hydra-concerts-tribute-professor-hans-tutschku

4.1 Act I, Scene 1: Discovery

The opening four-minute solo features the first dancer (the author) exploring the system's interactive vocabulary (Figure 5). Utilizing a music generation model fine-tuned on classical music, the mapping logic treats the stage as a large-scale interface where floor position correlates to pitch and eight-channel spatialization. Joint velocities map to MIDI velocity of the music, creating a legible link between physical effort and dynamic intensity. The scene concludes with a programmed visual cue—the dancer reaching behind the stage curtain—to invite the second performer to join the dance.

Figure 5: Dancer 1 performing solo in the first scene in front of the audience.

4.2 Act I, Scene 2: Connection

Upon the entry of the second dancer, a second musical voice joins the ongoing piano generation (Figure 6). In this three-minute sequence, each dancer controls a discrete musical voice based on their respective stage positions, continuing the timbral controls established in the solo. We introduce the spatial distance between the two performers as a new control parameter, contributing to timbre via distortion. The scene ends with a synchronized "collapse" gesture, which serve as a transition into a new genre.

Figure 6: Dancers explore the notion of proximity and how that influences their musical output via distortion

4.3 Act II, Scene 1: Exchange

The choreographic gesture to switch genres triggers a state change in the music generation engine from a model fine-tuned on classical music to a model fine-tuned on a few hours of jazz. To maintain rhythmic stability and more danceable music, the generation is forced to start with a drum foundation. When the dancers rise, they embody specific instruments (piano and trumpet) rather than individual voices, taking turns stepping into the spotlight.

The three-minute dialogue uses a spotlight spatial boundary where the system activates a dancer's assigned instrument only when they occupy the center-stage area. The performers engage in a structured call-and-response, where one dancer executes an eight-count phrase for the other to mirror. As both dancers move to share the spotlight, the Exchange evolves from discrete alternating solos into simultaneous movement. We utilize the ZED 2i's skeletal ID tracking to maintain instrument-to-person consistency, with a manual override available for potential ID swaps. The scene concludes with a hug, signaling the end of the dialogue.

4.4 Act II, Scene 2: Battle

The finale begins with a deceptive resolution, as the dancers exit the stage while the music fades to complete silence. The stillness suggests the work's conclusion, but the system abruptly re-initiates the generative process autonomously, prompting the dancers to return for a three-minute finale.

The freestyle expands from individual voices in the same genre to a Battle of styles—Jazz versus Disco—where each dancer embodies a distinct genre (Figure 7). To prevent complete chaos when playing these disparate genres simultaneously, both models use a shared drum loop to condition the generation. The finale investigates the boundaries of physical embodiment, pushing the performers to interpret the friction between their musical identities. As they alternate to take the spotlight, they showcase their specific genre before culminating in a chaotic battle where both performers must reconcile the friction of simultaneous, clashing genres at maximum physical energy. Due to hardware limitations of running two concurrent models with unlimited instruments each, we generated the tracks of the Battle offline to ensure performance stability.

5 Discussion

5.1 Dancer co-design experience

We begin our discussion with first-person reflections from both dancers. As a designer and co-performer using the system, Dancer 1 occupies a dual role as both the system's architect and its user. His reflection here adopts a reflexive practitioner perspective, acknowledging the inherent subjectivity of the experience as a source of deep, situated insight into the system's performance. In addition, Dancer 2, who performed with the system but had no technical involvement in the development of the system, expressed her views on interacting with the system in an interview.

5.1.1 System as instrument vs agent. The dancers' perception of the system shifted over the course of the performance, alternating between instrument-like and partner-like. In Act I, they experienced the system primarily as an instrument, mentally visualizing a piano-like interface mapped onto the performance space. Movement was associated with specific musical effects, though repeated gestures could yield different outcomes. Dancer 2 described this unpredictability as appealing but noted that it required an adjustment period, as shaping the music directly differed from her usual practice of reacting to a fixed musical source. In Act 2, the interaction felt less reactive and more atmospheric, with a stronger sense of being "inside" the sound rather than directly manipulating it. The dancers found this shift engaging,

(a) Dancer 1 freestyling **(b) Dancer 2 freestyling** **(c) Both dancers simultaneously**

Figure 7: Dancers exchanging and battling in the second act of the performance to music generation fine-tuned on jazz or disco.

noting that the contrast between instrument-like control and environmental immersion supported different movement qualities and modes of attention across sections.

5.1.2 Expectations of danceability. Historically, dance has functioned as a response to a fixed musical source so moving within an interactive environment felt unintuitive at first. Early rehearsals were difficult as the dancers navigated these abstract connections, eventually finding a flow by reflecting on the fundamental relationship between movement and sound. The piece was never performed in its entirety prior to the debut; while we established a structural road map of specific sections, the full, continuous interaction remained untested until the dancers were in front of a live audience. This was a deliberate choice to ensure that the social and musical negotiations were authentic rather than rehearsed. During the interaction, the dancers focused exclusively on each other rather than the audience to prioritize the piece as a private, embodied conversation mediated by AI than a traditional dance performance for observation.

While the system is perceived as more organic due to its variability and responsiveness, the unpredictability made it more difficult to choreograph precise musical accents. Sudden changes were described as more destabilizing in sparse musical textures, such as solo piano passages. In contrast, denser contexts with rhythmic grounding, including sections with drums and bass, allowed her to move with greater confidence, even when melodic output became unstable.

The interactions were not always seamless. At times, dancers felt their musicality was out of sync: the model might produce a sparse texture during a high-energy movement, or they might execute an accent into a musical silence. In traditional dance, this would be perceived as a mistake, yet with the generative framework, the author viewed these moments as a compelling aesthetic friction because they could not say who is right or wrong. These misalignments highlight the autonomy of the AI, forcing the dancers to constantly renegotiate movement in relation to a partner that was both responsive and unpredictable.

5.1.3 Rewarding moments of synchrony. The *Con Moto* system can produce structured, motif-driven, and danceable music. Moments when anticipated patterns returned and Dancer 1 was able to respond precisely felt especially rewarding:

> *The feeling of hitting the beat exists in conventional dance, but it carried a heightened significance here since the music had never existed before, was generated in real time, and emerged through our movement.*

In these instances, Dancer 1 experienced a strong sense of synchrony between their body and the system, as if they became one with the music.

5.1.4 Novel experiences with movement-sound relationship. Musicality is central to being a dancer; dancers learn to internalize musical form and predict upcoming accents to respond. Such is the case in known tracks, but Dancer 1 expressed that dancing with generated music pushed this concept to a new level:

> *"Performing with Con Moto demanded a higher degree of active listening than traditional dance sessions. Since the generative model is non-deterministic, we could not rely on pre-existing knowledge of the track's structure."*

Dancer 2's awareness of generating sound also directly shaped her choreographic decisions. She reported paying increased attention to verticality when height influenced musical parameters, incorporating spins to accentuate dynamic changes, and favoring arm gestures over leg-driven travel in order to maintain spatial stability for pitch control. These adjustments reflect an embodied negotiation between movement preference, spatial constraints, and perceived musical outcome. She also gravitated toward improvisational strategies over tightly structured movement, particularly in sections with denser musical textures.

5.1.5 Collaboration. The collaboration between the dancers allowed them to combine their internalized knowledge and tackle new interactions that emerged only through physical practice. Designing in proximity was challenging as the dancers navigated the uncertainty of the system to find the abstract connections that felt the most natural to their bodies. Throughout the performance, the dancers became different musical elements—shifting from embodying voices to instruments, and eventually representing different musical genres. Dancing with each other and with the system created a triadic bidirectional feedback loop on three levels: between each dancer and the system, and between the two performers themselves (Figure 8).

> *"While our movement led the music, and there would be no music without our movement, we also danced in response to music we generate by both ourselves and the other performer. No one was in total control; instead, the performance emerged from a resonance of the trio. We found that the more generative agency we afforded the system, the more the AI acted as a third body between us. We were not just dancing with a machine; we were dancing with each other through the machine. Con Moto did not just sonify*

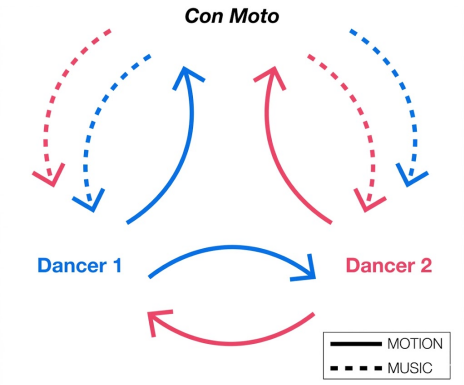

Figure 8: The triadic interaction topology of *Con Moto*. Solid arrows represent the flow of motion data from the performers to the system and each other, while dashed arrows indicate the generative musical feedback. The blue and red paths distinguish the individual influence and response cycles of each dancer within the collective loop.

> *our movements, it became a shared sonic space we inhabited together."*

5.2 Audience perception

We conducted a short voluntary audience survey following the performance to gauge the audience's perception of the dancer's interaction with the generated music. The surveyed audience members (n=6) do not have a technical understanding of the system employed in the performance.

Audience responses to the generated music suggest that perceptions of musical quality were closely tied to how the balance between dancer influence and musical autonomy was experienced during the performance. Many respondents described the music as immersive, present, and musically convincing, with several noting that they *"couldn't tell it was generated by AI."*

Sections that evoked a classical aesthetic were frequently highlighted as particularly effective, suggesting that stylistic coherence supported engagement even as the system responded to live movement.

At the same time, audience reflections indicate that the interaction was often perceived as neither fully dancer-led nor entirely music-led, but situated between these extremes. For some listeners, this balance contributed positively to the experience, reinforcing a sense of collaboration in which the music retained its own momentum while remaining responsive to the dancers. Others expressed a desire for either greater irregularity or clearer intervention, describing the music as "too perfect" and suggesting that increased unpredictability could enhance its naturalness. These differing reactions reflect varying expectations about how much agency the dancers should exert over the musical outcome.

Many audience members additionally expressed curiosity about how the system worked, suggesting an interest in the underlying mechanisms connecting movement and sound. The dancers particularly enjoyed the feedback:

> *"You looked very free."*

Overall, these responses suggest that audiences evaluated the generated music not only on its sonic qualities, but also on how convincingly the system navigated the balance between musical autonomy and dancer influence. Even when specific control relationships were not fully legible, the perception of a negotiated, responsive dynamic played a significant role in shaping how the music was received.

5.3 Technical limitations and future work

A primary limitation of the current system arises from generation latency and throughput. The generative model can only produce a limited number of tokens per second, which constrains the complexity of musical textures that can be generated in real time. While the classical solo piano model produces relatively coherent and expressive output at interactive rates, extending generation to full MIDI ensembles with more than three instruments remains challenging. In practice, the dance-oriented models, which include drums, bass, guitar, and melodic instruments, require substantial reduction in output density to maintain real-time responsiveness. This reduction can result in musical output that sounds overly simplified or "childish," highlighting the tension between ensemble richness and real-time performance constraints.

A second limitation concerns the current approach to enforcing spatial and ensemble-related constraints during generation. The present implementation applies negative masking to suppress unwanted pitch ranges or instruments, but this does not guarantee that the model will actively generate material in all desired regions. To compensate, the system inspects recent generation history and explicitly forces the next generated note to appear in a missing pitch range or instrument class when necessary. While effective in ensuring coverage, this strategy can introduce discontinuities, leading to incomplete or incoherent musical lines within a given register or instrument.

This issue is compounded by the structure of the symbolic representation, in which pitch and instrument class are determined after the onset time token is generated. As a result, the system may be forced to generate, for example, a percussive event in a spatial region where a melodic continuation would be more musically appropriate. Future work could address this limitation by incorporating multi-objective constraints directly into the generation process, enabling the model to reason jointly about pitch, instrument, and spatial distribution. Alternative directions include higher-throughput generation architectures, partial precomputation of ensemble stems, or planning-based approaches that enforce coverage and balance at a higher structural level rather than through reactive token-level interventions.

6 Conclusion

We have presented the research and co-design process of *Con Moto*, a real-time generative music system designed to facilitate human-AI partnership through dance improvisation. Our work culminated in a ten-minute live performance featuring two improvising dancers, followed by a detailed qualitative reflection on the resulting collaborative experience. The process of collaborating on a narrative performance balanced technical and artistic design considerations, resulting in a system with configurable agency. By moving beyond deterministic sonification toward a distributed, generative architecture, *Con Moto* explores the trade-offs between musical coherence and the low latency required for an organic dance experience.

Technically, we developed a system capable of generating multi-instrument MIDI sequences in different genres with real-time embodied steering of harmony, pitch range, and tempo, while simultaneously providing fine-grained embodied control

over velocity, timbre, and articulation. Throughout the process, the dancers' relationship with the system evolved from one of instrumental control toward an atmospheric partnership. Our findings suggest that resonance is not found in perfect synchrony, but rather in the aesthetic friction between the dancer's intent and the AI's generative autonomy. These moments of misalignment fostered a state of enhanced listening that transformed the performance into a private, embodied conversation between the two dancers and the AI system. While technical challenges remain regarding the throughput of complex multi-instrument ensembles, the current work demonstrates the viability of a generative creative medium through collaborative, real-time improvisation.

7 Ethical Standards

All performers and interview participants involved in this work provided informed consent for their participation and for the use of their feedback in research dissemination. No personal or identifying data beyond voluntary reflections were collected or reported.

The generative models used in this work were trained on publicly available MIDI datasets, which are known to be biased toward Western musical traditions due to the greater availability of freely accessible symbolic music in these genres. As a result, the musical styles produced by the system reflect these cultural biases. We acknowledge this limitation and view it as an important area for future work, particularly through the inclusion of more diverse musical corpora and performance practices as suitable datasets become available.

To mitigate environmental impact, we did not train large models from scratch and only fine-tuned pretrained models using comparatively lightweight training runs, each requiring approximately three hours of GPU time on NVIDIA RTX A6000 hardware.

Acknowledgments

The work was supported by MIT MGAIC Research Fund and the Quanta Computer Graduate Fellowship.

The authors would like to thank members of the MIT HAI-Res Lab, Hans Tutschku, and his "Composing with Max/MSP" class for their helpful feedback and discussions throughout the development of the system. We also thank Lancelot Blanchard for his assistance with the JAM_BOT system. Finally, we are grateful to Abena Kyereme-Tuah for performing as the dancer and for providing valuable feedback and insights into the system and its use in live performance.

References

[1] Gunjan Aggarwal and Devi Parikh. 2021. Dance2Music: Automatic Dance-driven Music Generation. arXiv:2107.06252 [cs.SD] https://arxiv.org/abs/2107.06252

[2] Tove Grimstad Bang and Sarah Fdili Alaoui. 2023. Suspended Circles: Soma Designing a Musical Instrument. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. Association for Computing Machinery, New York, NY, USA, 1–15. https://doi.org/10.1145/3544548.3581488

[3] Tove Grimstad Bang, Sarah Fdili Alaoui, and Elisabeth Schwartz. 2023. Designing in Conversation With Dance Practice. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. Association for Computing Machinery, New York, NY, USA, 1–16. https://doi.org/10.1145/3544548.3581543

[4] Frédéric Bevilacqua, Eric O Boyer, Jules Françoise, Olivier Houix, Patrick Susini, Agnès Roby-Brami, and Sylvain Hanneton. 2016. Sensori-Motor Learning with Movement Sonification: Perspectives from Recent Interdisciplinary Studies. *Frontiers in neuroscience*. 10 (2016), 385–. Place: Lausanne, Switzerland : Publisher: Frontiers Research Foundation.

[5] Lancelot Blanchard, Perry Naseck, Stephen Brade, Kimaya Lecamwasam, Jordan Rudess, Cheng-Zhi Anna Huang, and Joseph A. Paradiso. 2025. The jam_bot, a Real-Time System for Collaborative Free Improvisation with Music Language Models. In *Proceedings of the 26th International Society for Music Information Retrieval Conference (ISMIR 2025) (ISMIR)*. Daejeon, South Korea, 755–762. https://ismir2025program.ismir.net/poster_321.html

[6] Stephen Brade, Teng Ma, Lancelot Blanchard, Kimaya Lecamwasam, Carlos Mariano Salcedo, Suwan Kim, Perry Naseck, Andrew Li, Matthew R Michalek, Sebastian Franjou, and Anna Huang. 2026. Agents in Concert: A Case-Study of Bringing AI to the Stage in Practice. In *Proceedings of the 31st International Conference on Intelligent User Interfaces (IUI '26)*. Association for Computing Machinery, New York, NY, USA, 1340–1361. https://doi.org/10.1145/3742413.3789104

[7] Antoine Caillon and Philippe Esling. 2021. RAVE: A variational autoencoder for fast and high-quality neural audio synthesis. https://doi.org/10.48550/arXiv.2111.05011 arXiv:2111.05011 [cs].

[8] Ethan Chang, Zhixing Chen, Jb Labrune, and Marcelo Coelho. 2025. Be the Beat: AI-Powered Boombox for Music Suggestion from Freestyle Dance. In *Proceedings of the Nineteenth International Conference on Tangible, Embedded, and Embodied Interaction*. ACM, Bordeaux/Talence France, 1–6. https://doi.org/10.1145/3689050.3705995

[9] Sumanth Dathathri, Andrea Madotto, Janice Lan, Jan Hung, Ehsan Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. 2020. Plug and Play Language Models: A Simple Approach to Controlled Text Generation. In *International Conference on Learning Representations (ICLR)*. https://arxiv.org/abs/1912.02164

[10] Rebecca Fiebrink. 2017. Machine Learning as Meta-Instrument: Human-Machine Partnerships Shaping Expressive Instrumental Creation. In *Musical Instruments in the 21st Century*, Till Bovermann, Alberto De Campo, Hauke Egermann, Sarah-Indriyati Hardjowirogo, and Stefan Weinzierl (Eds.). Springer Singapore, Singapore, 137–151. https://doi.org/10.1007/978-981-10-2951-6_10

[11] Ian Hattwick, Joseph Malloch, and Marcelo Wanderley. 2014. Forming Shapes to Bodies: Design for Manufacturing in the Prosthetic Instruments. In *Proceedings of the International Conference on New Interfaces for Musical Expression*. Zenodo, 443–448. https://doi.org/10.5281/zenodo.1178792

[12] Curtis Hawthorne, Andriy Stasyuk, Adam Roberts, Ian Simon, Cheng-Zhi Anna Huang, Sander Dieleman, Erich Elsen, Jesse Engel, and Douglas Eck. 2019. Enabling Factorized Piano Music Modeling and Generation with the MAESTRO Dataset. arXiv:1810.12247 [cs.SD] https://arxiv.org/abs/1810.12247

[13] Thomas Hermann, Andy Hunt, and John G. Neuhoff (Eds.). 2011. *The Sonification Handbook*. Logos Verlag, Berlin. OCLC: ocn771999159.

[14] Kristina Hook. 2018. *Designing with the body : somaesthetic interaction design*. The MIT Press, Cambridge, Massachusetts.

[15] Takuma Kikuchi, Riki Saito, Risako Shibata, Atsuya Tsuchida, Kenshiro Taira, Nimisha Anand, Ryoho Kobayashi, Yuta Uozumi, and Shinya Fujii. 2025. transcriptions. In *Proceedings of the International Conference on New Interfaces for Musical Expression*, Sophie Rose, Jos Mulder, and Nicole Carroll (Eds.). Canberra, Australia, Article 34, 9 pages. https://doi.org/10.5281/zenodo.17801160 Live Performance.

[16] Ben Krause, Akhilesh Gotmare, Bryan McCann, Nitish Shirish Keskar, and Richard Socher. 2021. GeDi: Generative Discriminator Guided Sequence Generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. https://arxiv.org/abs/2009.06367

[17] Keon Ju Maverick Lee, Jeff Ens, Sara Adkins, Pedro Sarmento, Mathieu Barthet, and Philippe Pasquier. 2025. The GigaMIDI Dataset with Features for Expressive Music Performance Detection. *Transactions of the International Society for Music Information Retrieval* 8, 1 (2025), 1–19. https://doi.org/10.5334/tismir.203

[18] Qiaosheng Lyu and Ryo Ikeshiro. 2025. Between Garment and Prosthesis: The Design of an E-Textile Musical Interface. In *Proceedings of the International Conference on New Interfaces for Musical Expression*, Doga Cavdir and Florent Berthaut (Eds.). Canberra, Australia, Article 24, 4 pages. https://doi.org/10.5281/zenodo.15698825

[19] 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 (CC '25)*. Association for Computing Machinery, New York, NY, USA, 491–493. https://doi.org/10.1145/3698061.3734415

[20] Thomas Mitchell, Sebastian Madgwick, and Imogen Heap. 2012. Musical Interaction with Hand Posture and Orientation: A Toolbox of Gestural Control Mechanisms. In *Proceedings of the International Conference on New Interfaces for Musical Expression (NIME)*. https://www.nime.org/proceedings/2012/nime2012_44/

[21] 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*. ACM, Utrecht Netherlands, 1–9. https://doi.org/10.1145/3658852.3659072

[22] Teresa Pelinski, Andrew McPherson, and Rebecca Fiebrink. 2024. Ways of knowing, ways of writing: technical practice research in new musical instrument design. *Journal of New Music Research* 53, 1-2 (2024), 79–92. https://doi.org/10.1080/09298215.2024.2442348 arXiv:https://doi.org/10.1080/09298215.2024.2442348

[23] Andrew Pickering. 1995. *The mangle of practice : time, agency, and science*. University of Chicago Press, Chicago.

[24] Stefania Serafin, Stefano Trento, Francesco Grani, Hannah Perner-Wilson, Sebastian Madgwick, and Thomas Mitchell. 2014. Controlling Physically

Based Virtual Musical Instruments Using The Gloves. In *Proceedings of the International Conference on New Interfaces for Musical Expression (NIME)*. https://www.nime.org/proc/nime2014_258/

[25] John Thickstun, David Leo Wright Hall, Chris Donahue, and Percy Liang. 2024. Anticipatory Music Transformer. *Transactions on Machine Learning Research* (2024). https://openreview.net/forum?id=EBNJ33Fcrl

[26] Sam Trolland, Alon Ilsar, Ciaran Frame, Jon McCormack, and Elliott Wilson. 2022. AirSticks 2.0: Instrument Design for Expressive Gestural Interaction. In *Proceedings of the International Conference on New Interfaces for Musical Expression (NIME)*. https://doi.org/10.5281/zenodo.7293679

[27] Kevin Yang and Dan Klein. 2021. FUDGE: Controlled Text Generation with Future Discriminators. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*. https://arxiv.org/abs/2104.05218

[28] Fuming You, Minghui Fang, Li Tang, Rongjie Huang, Yongqi Wang, and Zhou Zhao. 2024. MoMu-Diffusion: On Learning Long-Term Motion-Music Synchronization and Correspondence. arXiv:2411.01805 [cs.SD] https://arxiv.org/abs/2411.01805

[29] Eevee Zayas-Garin and Andrew McPherson. 2022. Dialogic Design of Accessible Digital Musical Instruments: Investigating Performer Experience. *International Conference on New Interfaces for Musical Expression* (jun 16 2022). https://nime.pubpub.org/pub/dialogicadmis.

[30] Chaoyang Zhang and Yan Hua. 2024. Dance2Music-Diffusion: leveraging latent diffusion models for music generation from dance videos. *EURASIP Journal on Audio, Speech, and Music Processing* 2024, 1 (Sept. 2024), 48. https://doi.org/10.1186/s13636-024-00370-6