Human-in-the-Loop: Crossmodal AI Alignment between Movement and Audio Latent Spaces for Expressive Sonification in Dance Performance

Koray Tahiroğlu
Aalto University
School of Art, Design and Architecture
FI 00076 AALTO, Finland
koray.tahiroglu@aalto.fi

Mikael Hokkanen
Aalto University
School of Science
FI 00076 AALTO, Finland
mikael.hokkanen@aalto.fi

Ariana Marta
Aalto University
School of Art, Design and Architecture
FI 00076 AALTO, Finland
ariana.marta@aalto.fi

Abstract

This paper presents Crossmodal AI alignment, a generative AI framework for expressive sonification of human movement in dance performance. The system connects two variational autoencoders: a Movement VAE encoder, capturing real-time expressive dance movement features, and an Audio VAE (RAVE) decoder, generating corresponding musical textures and responses. A central alignment module links their latent spaces, allowing dynamic adaptation between movement and sound. Unlike fixed or rule-based mapping approaches, SonicMove alignment system introduces a human-in-the-loop alignment process, where the dancer calibrates the crossmodal relationship through embodied exploration prior to performance. This enables an adaptive and intuitive co-creative dialogue between performer and AI, producing sonifications that respond to subtle variations in movement. Exploratory sessions with invited dancers, centred on the latent space alignment process, suggest how performer-driven calibration shapes the perceived coherence and expressivity of generated sound in relation to movement, offering a possible direction toward more adaptive, multimodal performance systems that integrate movement, sound, and creative interpretation.

Keywords

NIME, Sonification, Human-in-the-loop, VAE, Alignment, Generative AI

1 Introduction

The sonification of human movement has long been explored within dance performances, interactive music systems, and embodied interaction research [5, 15, 17, 19, 21]. Many of these systems rely on motion capture or wearable sensors to extract kinematic features such as position, velocity, acceleration, joint angles, which are then mapped to parameters of sound synthesis. In most cases, this mapping is implemented through fixed functions, rule based systems, or procedurally defined parameter spaces. Rule-based or fixed mapping strategies tend to impose a relatively rigid relationship between movement and sound. The system often limits its ability to adapt to individual performers or to subtle changes in movement quality that occur during live performance. Once the mapping is structured, the space of possible interactions is predetermined. Reconfiguration of mappings usually requires code adjustments for a new mapping procedure, making it difficult for dancers to reshape the sonic behaviour of the system on their own.

At the same time, deep learning models have opened up new possibilities for working with the latent structure of both audio and human movement datasets. Variational autoencoders (VAEs) and related architectures have been successfully applied to learn low-dimensional latent spaces for music, timbre, and sonic textures [3], and similarly for pose sequences and movement dynamics [22]. These models provide continuous spaces where small changes in latent coordinates correspond to gradual and perceptually meaningful changes in the generated output. In NIME performance contexts, such latent spaces are generally explored with external controllers, such as sliders and touch interfaces, or simple mappings from sensor data [25, 30, 35]. This practice is effective to a point, but it still leaves open the question of how to tightly couple the latent structures of movement and sound in a way that feels expressive rather than only functional.

The work presented in this paper is motivated by the need for adaptive, generative methods that can support a more intuitive and co-creative dialogue between dancer and AI. The aim is to design a system in which the internal representations of movement and audio can be brought into correspondence, so that the expressive structure of one modality can be meaningfully reflected in the other. It is important to note here that we do not claim that there is a unique or universal alignment between movement and sound latent spaces. Instead, we present the SonicMove alignment system as a site of artistic practice and negotiation, shaped by the dancer's embodied exploration. To support this, our alignment module incorporates a human-in-the-loop process in which the dancer actively calibrates the crossmodal relationship through iterative experimentation and feedback.

2 Related Work

Generative AI models such as Variational Autoencoders (VAEs) and Generative Adversarial Networks (GANs) have become central tools for modelling complex data distributions and synthesising realistic new samples [8, 16, 27, 34]. By learning compact latent representations, these models support expressive forms of generation and transformation in high dimensional domains such as images, audio, and text [9, 14, 20, 28, 29]. Building on these developments in generative AI models, a substantial body of research investigates the relationship between music and bodily movement, particularly in the generation and sonification of dance movements. [11, 23]. Early approaches generate choreography from music using paired audio and dance datasets. Lee et al. [22], for example, use encoder-decoder architectures based on VAEs and GANs to generate 2D dance motion, while Tang et

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

al. [33] use LSTM autoencoders to map acoustic features to 3D motion capture data. These methods often struggle with diversity and stylistic specificity. Music2Dance [36] addresses some of these limitations with an autoregressive model using temporal convolutions, where higher level musical features such as rhythm, melody, and style act as control signals. This shift from low-level descriptors such as MFCCs or STFT magnitudes suggests that beat and rhythm are more fundamental to dance generation than generic spectral characteristics, highlighting the importance of musically meaningful intermediate representations when working across movement and sound.

Work with wearable sensors and motion capture has shown how real-time sonification can enhance dancers' awareness of their own movement and of spatial relationships in the performance space [13, 21]. Dancers in such studies often describe the systems as reciprocal, with sound influencing movement exploration and movement in turn, reshaping sound. These findings underline the importance of human feedback and subjective experience, motivating approaches in which performers actively shape the behaviour of the system. Human-in-the-Loop (HITL) learning has therefore emerged as a natural framework for such interactive and subjective domains. At the same time, HITL has rarely been applied to music generation. To our knowledge, Justus's work [18] is one of the few studies that explicitly investigate HITL reinforcement learning for music, using user ratings as rewards to steer generation towards individual preferences. Although promising, the authors note limitations related to large state spaces and relatively simple learning architectures. They suggest future directions involving neural network based models, alternative reward formulations such as pairwise preferences, and richer generative back ends including GANs. Generative AI models provide expressive yet compact representations for both audio and motion, while dance sonification research emphasises the need for perceptually grounded and responsive mappings. HITL methodologies further highlight that such mappings cannot be predefined, but must instead be shaped through human feedback. Informed by these insights, our work proposes an alignment model that integrates VAE latent representations with HITL guided learning to adapt movement–sound relationships to individual dancers.

3 SonicMove Alignment System Architecture

The alignment system comprises two variational autoencoders; one for movement and one for audio [1]. Both are trained independently on their respective modalities. Each VAE uses causal convolutions with zero look-ahead to respect the temporal structure of streaming input [4]. After training, the inference pathway comprises the movement encoder, an alignment module, and the audio decoder: the movement encoder maps body motion to a latent representation, which the alignment module projects to serve as the latent input for the audio decoder that renders time-continuous audio. The alignment module performs latent dimensionality mapping from movement to audio latent space, and up/downsampling of latent trajectories to match movement and audio rates. At runtime, both models employ inference-time caching to support continuous generation[4].

3.1 Movement VAE Encoder

The primary objective of the movement VAE is to learn a compact and well-behaved latent space in which continuous trajectories from real-time human motion serve as control signals for audio generation. To satisfy low-latency constraints, the encoder is deliberately kept lightweight in order to allocate more compute to the audio generation component. Three dancers participated in the data collection sessions. Data were recorded in two separate sessions, resulting in approximately 3.5 hours of motion data per dancer ($\approx$ 10.5 hours total). Each dancer wore three Xsens[26] inertial motion trackers attached to the right hand, chest, and right leg. The sensors were sampled at 60 Hz. For model training, ten features were extracted per sensor; linear velocity (x, y, z), linear acceleration (x, y, z), and orientation represented as unit quaternion (q_0, q_1, q_2, q_3). No normalisation or post-processing was applied; preliminary experiments showed that the raw measurements yielded a latent space with desired properties. With three sensors per dancer, the resulting input dimensionality was 30 features per time step at 60 Hz.

3.2 Audio VAE Decoder (RAVE)

The audio decoder generates continuous, sonically meaningful raw waveforms in real time by decoding latent trajectories. We employ RAVE (Realtime Audio Variational autoEncoder) [3], which uses a two-stage training process; first learning a compact latent space, then refining output quality with an adversarial objective. Prior work [7] demonstrates faster than real-time synthesis from a compact latent representation, which enables direct sampling and streaming capabilities.

3.3 Latent Space Alignment Module

The alignment module connects two pre-trained models: a Movement VAE encoder and an Audio VAE decoder [2]. Each of these learns a compact latent representation for its own modality, but the resulting latent spaces differ in dimensionality, temporal resolution, and structure. The alignment module learns to map movement latents to audio latents so that expressive gestures can be translated into sounds that feel coherent and musically meaningful. This alignment is implemented as a neural network trained with Human-in-the-Loop (HITL) feedback, allowing the dancer to iteratively refine the mapping based on how the audio is perceived in practice.

The alignment module consists of a lightweight feedforward neural network with a single linear layer followed by a ReLU activation function. The network takes an 8 dimensional movement latent vector as input and outputs a 4 dimensional audio latent vector. This simple architecture was chosen to prioritise stability and low computational cost, which are critical for real-time interaction. More complex architectures are considered as future extensions once the core alignment behaviour is validated.

During training, the system receives movement data from external sensors via OSC. The incoming data is collected over short temporal windows and stored in a buffer then passed through the Movement VAE encoder. The encoder outputs a latent representation z_{in}, which is intended to capture key characteristics of movement in that window. This vector is then fed to the alignment module, which produces an output latent vector z_{out} in the audio latent space. Because movement and audio operate at different temporal resolutions (60 Hz for movement and 44.1 kHz for audio), a temporal mismatch appears between their latent representations. To address this, the output of the alignment module is interpolated to match the temporal requirements of the audio decoder. This interpolation step makes it possible to

[1] Project repository is available at https://version.aalto.fi/gitlab/sopi/sonicmove_vae

[2] Alignment module is available at https://tinyurl.com/36425wfx

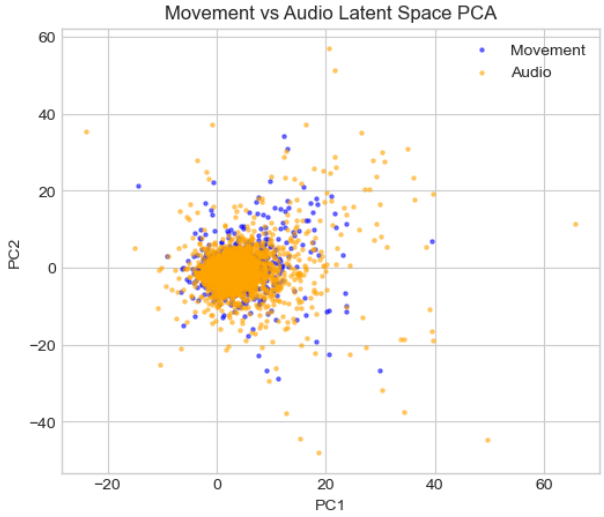

Figure 1: PCA projections of the movement and audio latent space.

generate a continuous audio latent trajectory from discrete movement inputs, supporting smooth sound synthesis and real-time performance.

In our initial experiment tests, directly decoding these aligned audio latent space often produced sounds that felt repetitive or too similar from one movement phrase to the next. Analysis of the audio latent space showed a high concentration of samples in dominant clusters, which made it quite likely that the decoded points would fall into the same region repeatedly. To encourage exploration of a wider range of perceptually interesting sounds, we introduced a centroid-based constraint mechanism. Prior to training, we computed a set of representative and musically interesting centroids from the audio latent space using clustering and stored them for later use.

The dancer can select or switch between these centroids, effectively choosing which region of the audio latent space to explore. After interpolation, the aligned latent vector is shifted toward the currently selected centroid before being decoded by the Audio VAE. This mechanism aims to balance stability and variety, allowing the dancer to move between distinct sonic behaviours while keeping the mapping coherent. Figure 1 illustrates the PCA projections of the movement and Figure 2 shows audio latent space, as well as the selected centroids used during training. After computing the final audio latent vector, it is decoded using the pre-trained Audio VAE decoder, and the resulting waveform is rendered in real time using the `sounddevice` library.

4 Human-in-the-Loop Alignment Process

The alignment process occurs before a performance, in dedicated sessions where the performers have time to personalise the system according to their own movement qualities, aesthetic preferences, and artistic intentions. Rather than relying on a predefined mapping, the system adapts through an iterative interaction in which the dancer's feedback directly guides the alignment model. The HITL training workflow runs through a series of cycles. At the beginning of a session, a centroid in the audio latent space is selected. The dancer then performs a movement sequence, captured in real time and stored in a buffer. The duration of each movement sequence is decided by the dancer, resulting in variable-length batches. Once the dancer ends the phrase, the buffered movement data is passed through the Movement VAE encoder, then through the alignment module, and finally through the Audio VAE decoder, producing an audio response corresponding to the captured movement. This generated sound is immediately played back, giving the dancer direct feedback on how the system has interpreted their motion.

After listening, the dancer provides a simple binary judgement indicating whether the result is acceptable or not. If the sound is approved, the alignment model is updated via backpropagation and optimisation. The loss is computed as the mean squared error between the input and output latent representations of the alignment module, (z_{in}, z_{out}). This reinforces the current mapping, making it more likely that similar movement inputs will reproduce similarly preferred audio responses in the future. If the sound is not approved, no update is applied. In this way, the model is prevented from learning from examples that do not align with the dancer's preferences. The process can be seen as a form of human-gated learning or preference-conditioned optimisation, where learning occurs only from positively evaluated examples. The alignment model functions as a transformation learner that is guided by human judgement.

Throughout training, the dancer can decide to keep the currently active centroid or switch to another one, which opens up different regions of the audio latent space. This encourages exploratory interaction and lets the performer investigate how different movement qualities might correspond to distinct sonic behaviours. The training session continues iteratively until the dancer chooses to stop. The alignment model is saved, so that training can be resumed later or a learnt mapping can be reused in performance.

5 Evaluation and Preliminary Findings

We evaluated the initial implementation of the SonicMove alignment system through exploratory, qualitative sessions with three invited dancers. The focus in these sessions was on the latent space alignment itself, not yet on a full real time performance setup. Each dancer participated in guided movement explorations while adjusting alignment parameters and commenting on how

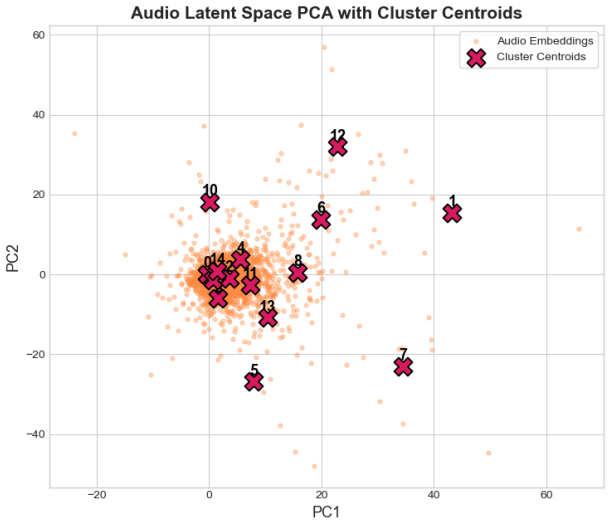

Figure 2: Audio Latent Space PCA with Cluster Centroids.

they experienced the relationship between their movement qualities and the generated sound.

5.1 Test Studies with Dancers

The exploratory test sessions took place over two consecutive days, Wednesday December 10th and Thursday December 11th 2025, in the multichannel studio at Marsio Studios, Aalto University. Each session lasted approximately three hours and focused on how the dancers engaged with the latent space alignment module. During the sessions, dancers wore Xsens inertial motion capture sensors and performed a series of guided movement tasks following a structured user-test study. Audio and video were recorded throughout to document the interaction process, system behaviour, and dancers' movement strategies, which later served as material for qualitative analysis and reflection.

Three professional dancers participated, all with extensive backgrounds in contemporary dance and artistic research contexts. Each was experienced in both improvisation and set choreography, in solo and ensemble formats, and all had worked with interactive or sound generating systems. After the introduction to the study, each dancer was guided through a structured set of four movement tasks designed to explore different expressive and temporal aspects of movement–sound relationships:

- **Sustained gestures**: continuous movements (such as slow arm arcs) emphasising smooth temporal evolution and continuity.
- **Micro-variations**: subtle changes in speed, weight, or shape to test sensitivity to fine grained expressive variation.
- **Accents**: clear, discrete gestures (for example, foot strikes or sudden arm snaps) performed on a steady pulse, focusing on responsiveness to rhythmic emphasis.
- **Dynamic contrast**: alternating between small and large movements and between slow and fast segments, highlighting changes in intensity and scale.

Each dancer completed all four tasks before the next participant began. On the first day, the system used a pretrained Darbouka RAVE v2 model (authored by Antoine Caillon) as the Audio VAE decoder, aligning the dancers' movement latent space with percussion-based audio generation. On the second day, the same movement tasks were repeated using a Vintage RAVE v1 model (also authored by Antoine Caillon), which allowed us to compare how interaction experiences shifted with different audio model characteristics.

During the study, after each task, dancers were asked to give a binary evaluation of the audio generated in relation to their movement, indicating with a simple yes or no whether they found the sonic response relevant or appealing. They were also invited to explain why they had chosen that judgement. These evaluation scores were fed back into the alignment module after each task, allowing the system to update its loss value incrementally. Following each evaluation, the system adjusted its internal alignment accordingly with the generated sonic responses. At the end of the second day, we conducted a semi structured group interview with all three dancers present. The participants responded in turn, reflecting on their own experiences, while also commenting on the observations of the other participants. The interview focused on perceptions of the alignment process; timing and responsiveness, the expressive qualities of the generated sound, and experiences of calibration and adaptation during the sessions. The aim was to gather reflective insight into how the alignment module shaped the movement–sound relationships, not to score performance quality or system accuracy. The interview was audio recorded and later transcribed for qualitative analysis.

5.2 Initial Thematic Coding

Following the semi-structured group interview, the recorded audio was transcribed using a verbatim-light approach to keep meaning while removing non-essential fillers. To analyse qualitative data, we followed a Thematic Analysis methodology, a widely used approach to identify meanungful patterns within qualitative datasets [2]. This approach allowed us to foreground the dancers' experiential accounts and still keeping our analytic steps reasonably transparent, which fits with the practice based character of the project. This approach allowed us to prioritise the dancers' experiential accounts while still keeping our analytic steps reasonably transparent, which fits well with the practice based nature of the project.

The analysis unfolded in two stages. In the first stage, we conducted initial coding to identify recurring ideas, concerns, and experiential qualities expressed by dancers in connection with the latent space alignment module. These codes were kept close to the participants' own wording and focused on how they perceived the relationships between movement and sound, and how they talked about agency during interaction. In the second stage, we grouped related codes into higher-level themes that captured shared patterns between participants. The themes that emerged *Perceived Coherence Between Movement and Sound, Embodiment and Bodily Thinking, Control vs. Surprise, Agency and Authorship* reflect key experiential dimensions of our AI alignment module within crossmodal generative setting.

5.2.1 Perceived Coherence Between Movement and Sound. Across all interviews, dancers consistently noted the perception of coherence between their movement and the generated sound. Dancers described a stronger sense of connection when the sonic output exhibited a clear structure, layering, and when it seemed to continue musically over time. They felt they could recognise traces of their own movement in what they were hearing.

This theme emerged from repeated comments about how clearly dancers could find connections with their movement. Their statements describing the ease of "detecting" movement in sound were coded under perceived coherence and movement–sound correspondence. Observations about richer layers or more "musical" sound materials were coded as musical structure and layering, while descriptions of disappointment when strong accents, pauses, or sudden changes did not appear in the sound were coded as mismatch or breakdown. These codes collectively formed the theme of perceived coherence, capturing how dancers evaluated the alignment between what they did and what the system produced.

> "I think I found more connections between my movement and sound today. Maybe also because I liked the sound more, so it was easier to connect to it. But also the sound material was maybe easier to approach — maybe it was closer to my preference as a dancer, so it was easier to relate to it."
> dancer #C

> "There were more layers and parameters in the sound, so it was actually more responsive. It was easier to detect your movement from the sound.

> Yesterday it felt more static or less organic, but today, because there was more structure in it, you could really hear when something changed. So that made it feel more connected."
> dancer #A

This perceived coherence appeared to be improved on the second day of testing, when dancers already had some experience with the system and when the audio model generated more layered and musically interesting textures. At those points, dancers noted that it was easier to build a relationship with the sound. Coherence broke down when the system failed to reflect key movement features, particularly accents or pauses. In such cases, dancers reported a sense of disconnection and described the movement–sound relationship as momentarily "falling apart."

5.2.2 Embodiment and Bodily Thinking. Dancers consistently talked about their interaction with the system through embodied reasoning, describing alignment as something understood through physical exploration rather than verbal explanation or abstract parameter adjustments. Sound preferences and evaluations were often described using bodily or movement metaphors, emphasising the body as the primary source of sense-making.

This theme emerged from codes that reflected how dancers understood and evaluated the system through bodily experience rather than verbal or abstract reasoning. Comments about "feeling" the relationship between movement and sound, or describing sound as an interpretation or extension of their bodily action, were grouped under embodiment and bodily metaphors. Descriptions of slowly getting to know the system by repeating movements, experimenting with variations, or observing how other dancers' gestures affected the sound were coded as learning through movement and observational learning. These comments point to a mode of engagement in which bodily thinking is the main way of working out what the system does and how to respond to it.

> "It feels more like it's an interpretation of the movement, like a sonic interpretation of what I did. We've done improvisations where one person dances and another person makes sounds as a response, like a gift. This felt similar. I do something, and then the AI gives me its interpretation of how it felt. It's not a mirror — it's more like a bodily translation."
> dancer #C

> "When you do constant movement and really let yourself follow it, you don't even remember exactly what you did. You just let the body go. And when there is some kind of match in the sound, it actually helps you recollect what the movement felt like. But you're also judging the sound at the same time, so it becomes this two-layer task; remembering the embodied experience and evaluating the sonic rendering."
> dancer #A

5.2.3 Control vs. Surprise. The tension between feeling in control and being surprised came up repeatedly in the interviews. On the second day, dancers described a clearer sense of agency and decision making, which they linked to a growing familiarity with both the system and the way the alignment behaved.

> "I had this feeling, I'm not sure if it was accurate, but when I was accelerating the movement, like going from slow to faster, I think it actually responded to some of those. It felt like it was picking up the change in speed sometimes. But I wasn't completely sure, because it didn't always do it. So there were moments where I thought, okay, now it's reacting — and other times when I expected something and nothing really happened."
> dancer #C

Surprise was often welcomed when it felt musically or expressively meaningful, for instance, when the system responded in ways that extended or reinterpreted movement qualities. However, surprise was experienced negatively when it took the form of a non-response or mistimed response, especially during accents, pauses, or sudden spatial changes. Dancers also noted latency at the start of sequences and the moments when responses felt temporally condensed, as if several events had been packed into a shorter span of time. These situations were described as breaking the feedback loop, shifting the experience from playful exploration to frustration.

> "With the accents it was so clear, it's not responding to pauses in any way. And then that feels disappointing. That sort of disrupts the connection.
> dancer #A

This theme brings together dancers' reflections on intentional control, unpredictability, and system responsiveness. Comments about increased confidence and clearer decision-making were coded as expressive control and learning and adaptation. Descriptions of unexpected sonic outcomes were coded as surprise, and further divided into musically meaningful surprise and negative surprise related to timing issues or contrast expectations, such as the absence of response to accents or pauses. The grouping these codes highlighted a recurring negotiation between control and unpredictability as a defining aspect of the interaction with the alignment module.

5.2.4 Agency and Authorship. Dancers expressed strong emotional reactions to moments where the system appeared to disregard their movement, when sonic responses were interrupted or failed to acknowledge the end of a movement phrase. These incidents were described not only as technical glitches, but also as moments where their sense of authorship and the value of their contribution as dancers felt challenged.

This theme emerged from codes related to dancers' sense of ownership, recognition and how dancers situated themselves within the interaction. Intense reactions to interruptions or ignored phrase endings were coded as emotional response and respect for movement phrasing. Comments asking for more influence over when sequences should start or stop were coded. These codes were grouped under agency and authorship, capturing the expectation that the system should acknowledge dancers as co creators whose expressive intentions matter.

> "At some point I stopped thinking of it as just reacting to me. It felt more like we were shaping it together. I would try something, then listen to what it did, and then adjust my movement again. It wasn't fully in my control, but it also wasn't random."
> dancer #C

Figure 3: The invited dancers participating in exploratory sessions, sharing their reflections and comments on the audio generated by the crossmodal alignment module. Their feedback helped shape the iterative calibration of movement and sound dynamics during testing.

> "I did this huge jump, really performative, and there was no response. That becomes disappointing. It feels like the system doesn't care about your movement. Especially when you expect something strong, like an explosion, and then nothing happens. That kind of surprise is not productive — it breaks the connection.
> dancer #A

Dancers expressed a clear desire to have control over when a sequence begins and ends, and to have their movement phrases recognised as complete expressive units. This expectation points toward a view of the system as a partner that should recognise dancers as co-authors and respond to their artistic decisions,

5.3 Observations and Results

Exploratory sessions with three invited dancers focused on interacting with and reflecting on the latent space alignment module (Figure 3). Across the two days, the dancers described how repeated engagement supported an embodied process of calibration, where they gradually developed an understanding of how their movement qualities were reflected in sound. This process was not described in terms of adjusting parameters, but as learning through bodily exploration and perceptual feedback over time.

Dancers repeatedly evaluated the system in terms of how clearly their movement was reflected in the generated sound. The perceived coherence increased once participants began to identify which movement characteristics were more legible to the system, particularly continuous dynamics, gradual transitions, and sustained gestures. One dancer commented, "I found more connections in my movement and in the sound" (Dancer C), while another noted that it became "more easy to detect your movement from the sound" (Dancer A). Coherence seemed to break down when key movement features, such as pauses, sharp accents, or sudden spatial changes, did not produce noticeable sonic changes that were described as perceptually mismatched.

Dancers explained their understanding of the alignment process primarily through embodied experience. They often used bodily metaphors and movement-based reasoning when evaluating the system's responses, describing sound as something to be "felt" or sensed through movement. One dancer characterised the generated sound as a kind of sonic interpretation of their movement, closer to another performer responding through sound (Dancer C). Watching how other dancers moved and how the system reacted to them also played a role; observational learning informed subsequent individual exploration and contributed to a shared sense of how the system behaved.

Dancers described interaction as an ongoing negotiation between control and surprise. On the second day of testing, dancers reported a stronger sense of intentional control, noting that it became "much easier… to say yes or no" to the system's responses (Dancer C). They appreciated forms of surprise that felt musically meaningful, where the system introduced variation while remaining perceptually connected to their movement. However, negative surprise emerged when the system failed to respond at expected moments, such as during accented gestures or pauses, or when initial latency at the beginning of a sequence disrupted continuity. These moments were often described as frustrating, not as creatively generative.

Perceived responsiveness was closely linked to dancers' sense of agency and authorship. As one dancer put it, when a strong movement produced no sonic response, it felt as though "the system doesn't care about your movement" (Dancer A). They expressed a wish for greater influence on when sequences started and stopped, highlighting the importance of respecting movement structure and closure in alignment-based sonification.

5.4 Discussion

These observations support the aims of the SonicMove alignment system, which seeks to support perceptually meaningful relationships rather than strict one-to-one mappings. The findings suggest that coherence is not merely a function of responsiveness but of whether the system captures structurally relevant aspects of movement in ways that performers can perceive and interpret musically. An increased sense of perceptual coherence was noted significantly when the movement unfolded through continuous dynamics and gradual transitions. This suggests that latent alignment may better capture higher-order movement qualities, such as flow, weight, and continuity, than discrete events or symbolic gestures.

The high level theme, Embodiment and Bodily Thinking, suggests that alignment operates as a form of bodily thinking in

which understanding emerges through action. This strongly supports the design rationale behind the human-in-the-loop alignment approach by allowing dancers to shape crossmodal relationships through embodied exploration. The system aligns with dancers' existing ways of thinking and learning through movement. A recurring theme in the dancers' accounts was the negotiation between control and surprise. While increased familiarity with alignment behaviour led to a stronger sense of agency, dancers also valued moments of sonic unpredictability when these were perceived as musically coherent. Importantly, surprise was welcomed when it extended or reinterpreted the movement but experienced negatively when it resulted from latency, truncation, or lack of response to key movements.

The strong reactions of the dancers to moments where the system appeared to ignore or override movement highlight the importance of perceived agency and authorship in co-creative systems. This suggests that agency in alignment-based systems is less about direct manipulation and more about being acknowledged within the interaction loop. The theme of Agency and Authorship directly supports the motivation for a human-in-the-loop alignment process. Here, agency emerges not from full control but from the ability to meaningfully influence the behaviour of the generative system. The binary evaluative feedback mechanism can be understood as a minimal but meaningful form of authorship. By allowing dancers to influence system behaviour through embodied judgment, the alignment module supported a form of co-creation grounded in performative evaluation.

Our approach resonates with Fdili Alaoui's work on movement qualities as interaction modality [1, 10], which emphasises capturing how movement is performed rather than what spatial trajectory it follows. Like Fdili Alaoui's systems for Laban Effort sonification [12], SonicMove focuses on qualitative aspects of movement, such as flow, weight, continuity as the basis for sonic generation. However, Fdili Alaoui's mapping-by-demonstration approach learns fixed movement-sound associations from prerecorded examples, but our latent space alignment module allows performers to iteratively recalibrate these associations through real-time feedback during dedicated training sessions. This positions the dancer not as a user of predetermined mappings but as an active co-designer of the system's interpretive behaviour.

The human-in-the-loop paradigm in the SonicMove alignment system addresses what Caramiaux and Fdili Alaoui [6] identify as a central friction in artistic AI practice, the tension between artistic agency and the black-box of algorithmic decision-making. Rather than approaching AI as either a transparent tool or an autonomous collaborator, SonicMove frames the alignment process as a negotiation, a space where the dancer's embodied judgment continuously reshapes how movement latents map onto sonic textures. This also resonates with broader discussions in HCI about agency and authorship in human-AI co-creation [24, 31], where ownership emerges not from total control but from meaningful influence over system behaviour.

The question of authorship becomes particularly important in generative AI contexts. When a dancer approves or rejects generated sounds during training, they are not simply tuning parameters, but they are negotiating aesthetic preferences that become inscribed into the alignment module's learned weights. This process distributes authorship across human judgment, machine learning dynamics, and the material properties of the latent spaces themselves. In this sense, our system resonates with the ways Barad's notion of intra-action has been disucssed in NIME context [32] where the dancer and the AI do not interact as pre-existing independent entities but co-constitute one another through their ongoing entanglement. The alignment module becomes a site where movement and sound are mutually shaped not by imposing one modality onto the other, by finding temporary alignments that respect the expressive specificities of both.

This approach, for the NIME field, suggests a design direction that moves beyond fixed mappings or fully autonomous generative agents. SonicMove demonstrates how latent space alignment, combined with performer-driven calibration, can support systems that are neither rigidly deterministic nor unpredictably black-box. Instead, SonicMove alignment is adaptive, learnable, and responsive to individual movement aesthetics while retaining enough consistency to be musically meaningful. Future work may explore how alignment mechanisms can better accommodate temporal articulation, spatial awareness, and performer-driven segmentation, further supporting the expressive needs of dancers and other movement-based practitioners.

6 Conclusion

In this paper, we presented our SonicMove alignment system that aligns movement and audio latent spaces through human-in-the-loop training, and suggests how such a setup can matter for embodied AI performance practices. Building alignment as something dancers negotiate through bodily exploration seems to open space for co-creation, where the AI does not fix the aesthetic outcome in advance but responds to evolving movement strategies, shifting senses of coherence and surprise, and ongoing questions of agency. These findings may be useful for choreographic research and for NIME performance contexts. In the near future, we will see several directions for extension, including alignment between audio latent spaces that concentrate on raw timbre and real-time timbral reshaping, as well as integration into modular NIME systems.

7 Ethical Standards

This research was supported by Business Finland funding within the authors' home organisation and did not receive external commercial sponsorship. The authors declare no financial or non-financial conflicts of interest. The user-test sessions with invited dancers were conducted under approval from Aalto Univertsity School of ARTS institutional ethics committee; all participants received written and verbal information about the study and gave informed consent prior to participating, including consent for audio and video recording.

Acknowledgments

This work was supported by SonicMove Project, which is a EU (NextGenerationEU-funding) and Business Finland 7655/31/2022 co-research project in collaboration with VTT, the University of Eastern Finland, Aalto University, Taustamarkkinat BGMT Oy, and Genelec Oy. We would also like to thank the three dancers who participated in our test studies for their valuable feedback and comments.

References

[1] Sarah Fdili Alaoui, Baptiste Caramiaux, Marcos Serrano, and Frédéric Bevilacqua. 2012. Movement qualities as interaction modality. In *Proceedings of the Designing Interactive Systems Conference*. 761–769.

[2] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative research in psychology* 3, 2 (2006), 77–101.

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

[4] Antoine Caillon and Philippe Esling. 2022. Streamable Neural Audio Synthesis With Non-Causal Convolutions. arXiv:2204.07064 [cs.SD] https://arxiv.org/abs/2204.07064
[5] Antonio Camurri, Barbara Mazzarino, and Gualtiero Volpe. 2003. Analysis of expressive gesture: The eyesweb expressive gesture processing library. In *International gesture workshop*. Springer, 460–467.
[6] Baptiste Caramiaux and Sarah Fdili Alaoui. 2022. " Explorers of Unknown Planets" Practices and Politics of Artificial Intelligence in Visual Arts. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2 (2022), 1–24.
[7] caspe franco, shier jordie, sandler mark, saitis charalampos, and mcpherson andrew. 2025. designing neural synthesizers for low-latency interaction. *journal of the audio engineering society* 73 (february 2025), 240–255. Issue 5.
[8] Hao-Wen Dong, Wen-Yi Hsiao, Li-Chia Yang, and Yi-Hsuan Yang. 2017. MuseGAN: Multi-track Sequential Generative Adversarial Networks for Symbolic Music Generation and Accompaniment. arXiv:1709.06298 [eess.AS] https://arxiv.org/abs/1709.06298
[9] Jesse Engel, Kumar Krishna Agrawal, Shuo Chen, Ishaan Gulrajani, Chris Donahue, and Adam Roberts. 2019. GANSynth: Adversarial Neural Audio Synthesis. arXiv:1902.08710 [cs.SD] https://arxiv.org/abs/1902.08710
[10] Sarah Fdili Alaoui, Frédéric Bevilacqua, Bertha Bermudez Pascual, and Christian Jacquemin. 2013. Dance interaction with physical model visuals based on movement qualities. *International Journal of Arts and Technology* 6, 4 (2013), 357–387.
[11] Jules Françoise, Norbert Schnell, and Frédéric Bevilacqua. 2013. A multimodal probabilistic model for gesture–based control of sound synthesis. In *Proceedings of the 21st ACM International Conference on Multimedia* (Barcelona, Spain) *(MM '13)*. Association for Computing Machinery, New York, NY, USA, 705–708. https://doi.org/10.1145/2502081.2502184
[12] Jules Françoise, Sarah Fdili Alaoui, Thecla Schiphorst, and Frédéric Bevilacqua. 2014. Vocalizing dance movement for interactive sonification of laban effort factors. In *Proceedings of the 2014 conference on Designing interactive systems*. 1079–1082.
[13] Andrea Giomi. 2020. Somatic Sonification in Dance Performances. From the Artistic to the Perceptual and Back. In *Proceedings of the 7th International Conference on Movement and Computing* (Jersey City/Virtual, NJ, USA) *(MOCO '20)*. Association for Computing Machinery, New York, NY, USA, Article 7, 8 pages. https://doi.org/10.1145/3401956.3404226
[14] 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
[15] Thomas Hermann, Oliver Höner, and Helge Ritter. 2005. AcouMotion–an interactive sonification system for acoustic motion control. In *International Gesture Workshop*. Springer, 312–323.
[16] Cheng-Zhi Anna Huang, Ashish Vaswani, Jakob Uszkoreit, Noam Shazeer, Ian Simon, Curtis Hawthorne, Andrew M. Dai, Matthew D. Hoffman, Monica Dinculescu, and Douglas Eck. 2018. Music Transformer. arXiv:1809.04281 [cs.LG] https://arxiv.org/abs/1809.04281
[17] Alexander Refsum Jensenius and Rolf Inge God. 2013. Sonifying the shape of human body motion using motiongrams. *Empirical Musicology Review* (2013), 73–83.
[18] Aju Ani Justus. 2025. Music Generation using Human-In-The-Loop Reinforcement Learning. arXiv:2501.15304 [cs.SD] https://arxiv.org/abs/2501.15304
[19] Ajay Kapur, George Tzanetakis, Naznin Virji-Babul, Ge Wang, and Perry R Cook. 2005. A framework for sonification of vicon motion capture data. In *Proceedings of the 8th International Conference on Digital Audio Effects (DAFX-05)*. 20–22.
[20] Jong Wook Kim, Rachel Bittner, Aparna Kumar, and Juan Pablo Bello. 2018. Neural Music Synthesis for Flexible Timbre Control. arXiv:1811.00223 [cs.SD] https://arxiv.org/abs/1811.00223
[21] Aleksandar Koruga and Koray Tahiroğlu. 2024. Dance Movement and Sound Cross-Correlation; Synthesis Parameters on the Micro and Meso Musical Time Scales. In *Proceedings of the 19th International Audio Mostly Conference: Explorations in Sonic Cultures* (Milan, Italy) *(AM '24)*. Association for Computing Machinery, New York, NY, USA, 445–456. https://doi.org/10.1145/3678299.3678361
[22] Juheon Lee, Seohyun Kim, and Kyogu Lee. 2018. Listen to Dance: Music-driven choreography generation using Autoregressive Encoder-Decoder Network. *ArXiv* abs/1811.00818 (2018). https://api.semanticscholar.org/CorpusID:53301150
[23] Leman M. Lesaffre M. et al Maes, PJ. 2009. From expressive gesture to sound. *J Multimodal User Interfaces 3, 67–78 (2010)* (2009).
[24] Landon Morrison and Andrew McPherson. 2024. Entangling entanglement: A diffractive dialogue on HCI and musical interactions. In *Proceedings of the 2024 CHI conference on human factors in computing systems*. 1–17.
[25] Tim Murray-Browne and Panagiotis Tigas. 2021. Latent Mappings: Generating Open-Ended Expressive Mappings Using Variational Autoencoders. arXiv:2106.08867 [cs.HC] https://arxiv.org/abs/2106.08867
[26] Monique Paulich, Martin Schepers, Nina Rudigkeit, and G. Bellusci. 2018. Xsens MTw Awinda: Miniature Wireless Inertial-Magnetic Motion Tracker for Highly Accurate 3D Kinematic Applications. https://doi.org/10.13140/RG.2.2.23576.49929
[27] Adam Roberts, Jesse Engel, Colin Raffel, Curtis Hawthorne, and Douglas Eck. 2019. A Hierarchical Latent Vector Model for Learning Long-Term Structure in Music. arXiv:1803.05428 [cs.LG] https://arxiv.org/abs/1803.05428
[28] Tarren Sexton. 2023. MuseNet. *Music Reference Services Quarterly* 26, 3-4 (2023), 151–153. https://doi.org/10.1080/10588167.2023.2247289 arXiv:https://doi.org/10.1080/10588167.2023.2247289
[29] Koray Tahiroğlu, Miranda Kastemaa, and Oskar Koli. 2021. Ganspacesynth: A hybrid generative adversarial network architecture for organising the latent space using a dimensionality reduction for real-time audio synthesis. In *Conference on AI Music Creativity*.
[30] Koray Tahiroglu and Lonce Wyse. 2024. Latent Spaces as Platforms for Sonic Creativity. In *Proceedings of the 16th International Conference on Computational Creativity, ICCC*, Vol. 24.
[31] Koray Tahiroğlu. 2021. Ever-shifting roles in building, composing and performing with digital musical instruments. *Journal of New Music Research* 50, 2 (2021), 155–164. https://doi.org/10.1080/09298215.2021.1900275 arXiv:https://doi.org/10.1080/09298215.2021.1900275
[32] Koray Tahiroğlu. 2024. Musical intra-actions with digital musical instruments. *Journal of New Music Research* 53, 1-2 (2024), 126–138. https://doi.org/10.1080/09298215.2024.2442350 arXiv:https://doi.org/10.1080/09298215.2024.2442350
[33] Taoran Tang, Jia Jia, and Hanyang Mao. 2018. Dance with Melody: An LSTM-autoencoder Approach to Music-oriented Dance Synthesis. In *Proceedings of the 26th ACM International Conference on Multimedia* (Seoul, Republic of Korea) *(MM '18)*. Association for Computing Machinery, New York, NY, USA, 1598–1606. https://doi.org/10.1145/3240508.3240526
[34] Li-Chia Yang, Szu-Yu Chou, and Yi-Hsuan Yang. 2017. MidiNet: A Convolutional Generative Adversarial Network for Symbolic-domain Music Generation. arXiv:1703.10847 [cs.SD] https://arxiv.org/abs/1703.10847
[35] Shuoyang Zheng, Anna Xambó Sedó, and Nick Bryan-Kinns. 2024. A Mapping Strategy for Interacting with Latent Audio Synthesis Using Artistic Materials. arXiv:2407.04379 [cs.SD] https://arxiv.org/abs/2407.04379
[36] Wenlin Zhuang, Congyi Wang, Siyu Xia, Jinxiang Chai, and Yangang Wang. 2020. Music2Dance: DanceNet for Music-driven Dance Generation. arXiv:2002.03761 [cs.CV] https://arxiv.org/abs/2002.03761

A Online Resources

A set of recorded video materials from the dancer user studies has been made available online. For each dancer, there are two videos corresponding to each day of the exploratory sessions. In these videos, the dancers perform a structured set of four movement tasks. After each task, they provide a binary evaluation of the resulting generated audio, and offering their comments on how they experienced the relationship between their movement qualities and the sound produced by the system.

DAY 1:
Dancer A - https://vimeo.com/1161445299
Dancer B - https://vimeo.com/1161442583
Dancer B - https://vimeo.com/1161447232

DAY 2:
Dancer A - https://vimeo.com/1161449685
Dancer B - https://vimeo.com/1161448301
Dancer B - https://vimeo.com/1163722775