Reimagining Dance: Real-time Music Co-creation between Dancers and AI

Olga Vechtomova[*] and **Jeff Bos**[**]

[*]University of Waterloo, Canada
[**]WordSynth Inc.
ovechtom@uwaterloo.ca, jeff@wordsynth.com

Abstract

Dance performance traditionally follows a unidirectional relationship where movement responds to music. While AI has advanced in various creative domains, its application in dance has primarily focused on generating choreography from musical input. We present a system that enables dancers to dynamically shape musical environments through their movements. Our multi-modal architecture creates a coherent musical composition by intelligently combining pre-recorded musical clips in response to dance movements, establishing a bidirectional creative partnership where dancers function as both performers and composers. Through correlation analysis of performance data, we demonstrate emergent communication patterns between movement qualities and audio features. This approach reconceptualizes the role of AI in performing arts—as a responsive collaborator that expands possibilities for both professional dance performance and improvisational artistic expression across broader populations.

Introduction

Dance and music traditionally exist in a hierarchical relationship where movement follows sound. Typically, choreographers design dance to existing music, or collaborate with composers to create accompanying scores. Even in improvisational dance, performers respond to pre-composed or live music, but rarely influence the musical composition itself.

Artificial intelligence now offers an opportunity to invert this relationship. While most AI systems in dance maintain the traditional paradigm by generating choreography from musical input, our research proposes a fundamental shift: enabling dancers to dynamically shape musical environments through their movements. This approach reconceptualizes dancers as both performers and composers, establishing a bidirectional creative partnership between human movement and AI-created sound.

In this paper, we present a system that enables dancers to dynamically influence musical composition in real-time through their movements. Our technical contribution is a multi-modal architecture (Figure 1) that selects and seamlessly combines pre-recorded musical clips in response to a dancer's movement patterns. Unlike previous approaches, our system creates an evolving musical environment where the dancer becomes an active co-creator of the soundscape rather than merely responding to it. We demonstrate through correlation analysis of pilot performances that this approach creates emergent communication patterns between dancer and system, establishing meaningful bidirectional relationships between specific movement qualities and audio features.

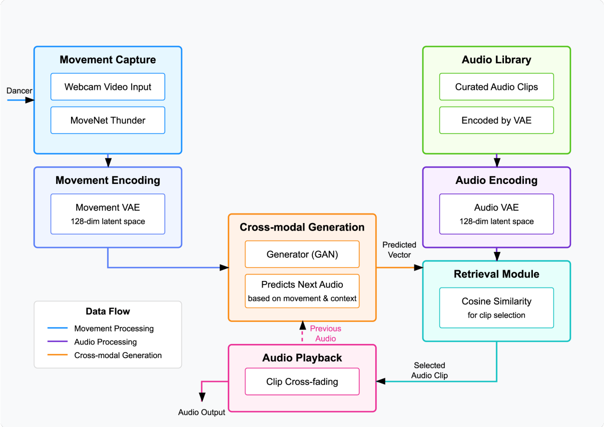


Figure 1: System diagram

This approach not only transforms dance performance but also opens new theoretical and practical directions for studying creative partnership between human performers and AI systems. While considerable research has examined AI's role in generating static artistic content like images or poetry, the dynamic, real-time nature of dance performance presents unique challenges that remain largely unexplored.

Related Work

Existing research in dance and AI has primarily focused on generating choreographic movements from musical input. Tang et al. (2018) developed an LSTM-autoencoder model that synthesizes dance choreography by mapping acoustic features to motion features, addressing the challenge of selecting appropriate dance figures that match musical elements. Lee et al. (2019) proposed a synthesis-by-analysis framework that decomposes dance into basic units to generate style-consistent and beat-matching movements from music. While these approaches demonstrate technical sophistication in movement generation, they maintain the traditional unidirectional relationship where dance follows music.

Some researchers have begun exploring more interactive approaches to AI in dance. Kumar et al. (2020) developed LuminAI, an improvisational dance installation where an AI agent dances with users, implementing a lead-and-follow dynamic based on creativity metrics. The choreographic duo AΦE's recent "Lilith.Aeon" performance, as reported in The Guardian (Winship 2024), demonstrated how an AI system trained on human-generated movements could become an active creative partner, suggesting new movement possibilities while maintaining the distinctive style of the choreographers. The Royal Ballet choreographer Wayne McGregor's collaboration with Google Arts & Culture Lab produced AISOMA, a system that suggests new movement variations by analyzing rehearsal videos, expanding the choreographic possibilities available to dancers and choreographers (Winship 2024).

Theoretical Foundations. This research builds on several theoretical foundations that help frame our understanding of human-AI creative collaboration. Particularly relevant is the concept of "mixed-initiative creative interfaces" developed by Deterding et al. (2017), which describes systems where human and computational agents take turns contributing to an evolving artistic work. In the context of dance performance, this framework takes on new dimensions as the collaboration happens in real-time, with the dancer's physical movements and the AI's musical responses creating a dynamic feedback loop.

Additionally, our work is informed by computational creativity concepts such as Colton and Wiggins' (2012) ”creative responsibility,” where AI systems take on creative roles beyond mere tools—evaluating aesthetics and inventing processes. This complements Jennings' (2010) notion of "creative autonomy," which requires systems to independently evaluate and evolve their standards.

While the pilot study reported in our paper establishes technical foundations, these frameworks guide our vision for AI systems that can function as genuine creative partners in dance performance.

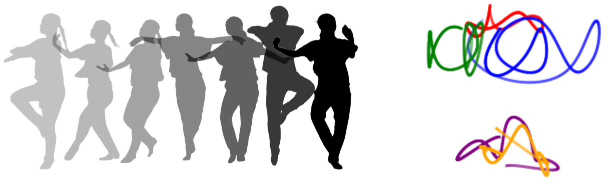
Beyond computational creativity, performance studies literature on improvisation and real-time creative decision-making provides another crucial theoretical perspective. Foundational work by Bailey (1992) and Nachmanovitch (1990) established key principles of improvisational practice, while recent scholarship addresses the complexities of dance improvisation and human-machine interaction. De Spain's (2014) topographical approach illuminates how dancers make moment-to-moment decisions, and Foster (2002) examines how improvisational structures emerge through real-time choreographic choices. For human-AI creative collaboration specifically, Hoffman and Weinberg's (2011) work on interactive robotic improvisation offers frameworks for understanding how performers adapt to non-human partners. Carter's (2000) analysis of improvisation as breaking established conventions to discover new artistic expressions that ”could not be found in a systematic preconceived process” is particularly relevant. Following Carter, our system aims to create new paradigms that enable real-time invention and discovery through the act of creation itself.

System Architecture

Our system enables real-time generation of responsive musical accompaniment to dance movements through a multi-stage machine learning pipeline. The architecture comprises three primary components: (1) an audio encoding/decoding system, (2) a movement encoding system, and (3) a cross-modal generation network that predicts appropriate musical responses to movement. These components work in concert to create a cohesive interactive performance environment.

Audio Representation Learning. To learn audio representations, we trained a Variational Autoencoder (VAE) (Kingma and Welling 2014) on spectrograms of 3.5-second audio clips. The audio VAE consists of convolutional layers for both encoding and decoding, with a 128-dimensional latent space representation. This architecture effectively compresses spectrograms into a compact latent code that preserves meaningful acoustic properties while discarding noise. The encoder uses five convolutional layers with ReLU activations to transform input spectrograms (224×224×1) into a latent distribution, while the decoder reverses this process through transposed convolutions.

Movement Representation Learning. To encode dance movements, we developed a parallel VAE architecture that processes visual representations of movement trajectories. The movement data is collected by using TensorFlow MoveNet Thunder pipeline, which analyzes either pre-recorded videos during training or webcam video stream during inference. Rather than working with raw skeletal joint data, we first transform movement sequences into color-coded trajectory images (Figure 2b). Each 3.5-second movement sequence is represented as an RGB image (256×256×3) where five key landmarks (head, left wrist, right wrist, left ankle, right ankle) are visualized as coloured curves showing their trajectories over time.



(a) Dance timelapse photo

(b) Movement trajectories

Figure 2: (a) Timelapse photo of dancer's movements. (b) Trajectories of dancer's movements captured by our system. The landmarks are colour-coded as follows: red - head, green - left wrist, blue - right wrist, orange - left ankle, magenta - right ankle.

This image-based approach allows us to leverage convolutional architectures commonly used in computer vision while capturing the temporal dynamics of movement. The movement VAE employs a structure parallel to the audio VAE, with the encoder producing a 128-dimensional latent vector that encapsulates the essential spatial and temporal characteristics of the dance movement.

Cross-Modal Generation The Generative Adversarial Network (GAN) (Goodfellow et al. 2014) bridges the movement and audio domains. The GAN's generator takes two inputs: (1) the latent representation of the current movement and (2) the latent representation of the previous audio clip. It then predicts the latent vector for the next audio clip that would best complement the current movement.

The generator employs a latent combiner module that integrates movement and audio latent vectors. While we experimented with several combination methods (concatenation, multiplication, and various learned approaches including gated, FiLM, and cross-attention mechanisms), we found that pointwise addition produces the most effective results. This addition operation is followed by a multi-layer network with hidden dimensions of 256 units, LayerNorm for stabilization, and LeakyReLU activations.

Retrieval Module Rather than directly decoding the predicted latent vector, which could result in lower audio quality, we employ a retrieval-based approach. We calculate the cosine similarity between the predicted latent vector and the latent representations of clips in our reference database. The audio clip with the highest similarity is selected, ensuring high-quality output while maintaining contextual relevance.

Figure 3 shows how both previous audio and movement influence the system's predictions. With identical previous audio but different movements, the system selects dramatically different clips: ambient music for minimal movement (3a) versus rhythmic clips for energetic break-dancing (3c). Similarly, when movement remains constant but previous audio changes (3a vs. 3b), the predicted clips also differ.

Real-time Inference We developed a React/NodeJS application, which operates as follows during live performance:

Movement Capture: A webcam captures the dancer's movements, which are processed by a pose estimation model (TensorFlow MoveNet Thunder) to extract the five key landmarks.

Movement Encoding: The landmarks' trajectories are rendered as a color-coded image and encoded by the movement VAE into a latent vector.

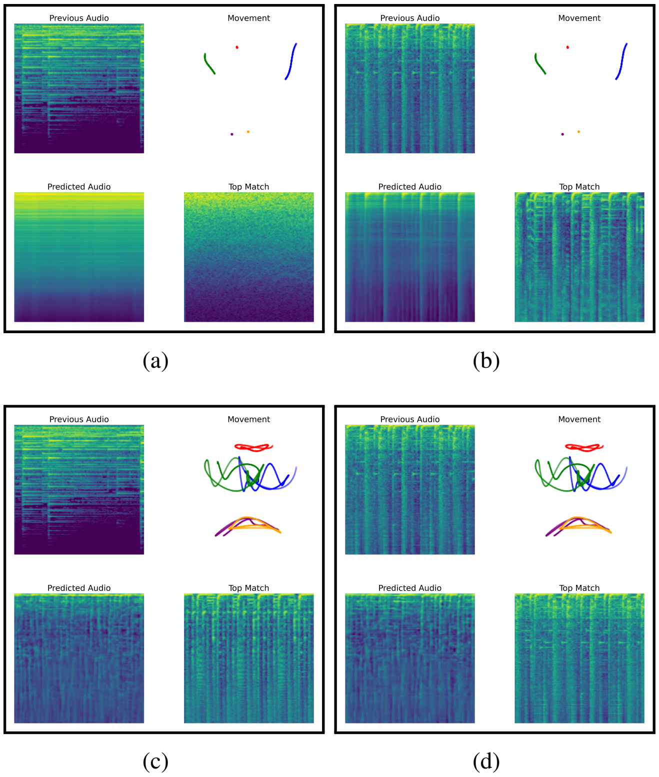
Audio Prediction: The movement latent vector and the previous audio clip's latent vector are fed into the GAN, which predicts the next audio clip's latent representation.

Clip Selection: The system retrieves the audio clip whose latent representation has the highest cosine similarity to the predicted vector.

Audio Playback: The selected clip is cross-faded with the currently playing audio to create seamless transitions.

The dancer can also curate the source audio library from which the system selects clips, allowing performers to influence the overall sonic palette and musical style of their movement-conditioned compositions.

Dataset. The aligned video-audio dataset used to train the movement VAE and GAN consists of 18K 3.5-second recordings, 17K of which were sourced from the AIST dance dataset (Tsuchida et al. 2019) and 1K from our own dataset containing video material recorded specifically for this project or provided to us by professional dance collaborators. The audio VAE was trained on a larger set of 50K audio clips from the authors' studio recordings, spanning multiple genres of electro-acoustic music with varying tempos and instrumentation.



(a) (b)

(c) (d)

Figure 3: System response examples showing the influence of both inputs on prediction.

Pilot study

We conducted a pilot study with three participants of varying dance experience: P1 (10+ years ballet), P2 (2-3 years ballet/jazz), and P3 (no formal training). Each dancer performed improvisational movement with the system for up to 30 minutes. We recorded both video and system-generated audio, collecting over 70 minutes of data. Our analysis aimed to identify relationships between dance movements and generated audio, particularly examining which audio features correlate most strongly with movement intensity.

Video and Audio Features. We segmented performances into 10-second clips and extracted movement data using MoveNet Thunder to track key body points (head, wrists, ankles) normalized to [-1, 1]. Movement energy was measured as the Euclidean distance between corresponding points in consecutive frames, with statistical measures (mean, min, max, standard deviation) computed for each clip. For audio, we extracted 47 features using Librosa and Essentia, including spectral features (MFCCs, contrast, flux), chroma features, and psychoacoustic measures. Key audio features in our analysis include *MFCCs* (representing timbre, with *mfcc_1* capturing overall spectral shape), *spectral contrast* (peak-valley differences across frequency bands), *chroma* (pitch class distribution), and *spectral flux* (frame-to-frame spectral changes).

Statistical Analysis[1] To explore the relationship between dance movement and system-generated audio, we applied several statistical methods. Pearson correlation analysis was performed to assess the linear relationships between individual movement energy measures (average, maximum, minimum, and standard deviation) and audio features. Principal Component Analysis (PCA) was conducted separately on video and audio features to identify patterns of variance and reduce dimensionality, while Canonical Correlation Analysis (CCA) was used to examine the multivariate relationships between movement and audio feature spaces.

Additionally, we used Partial Least Squares (PLS) regression to model predictive relationships between the two modalities, evaluating how well sets of audio features could predict movement energy metrics, and vice versa. To identify the most influential audio features, we employed Random Forest regression models, computing feature importance scores based on their contribution to predicting each movement energy statistic. The quality of these predictive models was evaluated using the coefficient of determination (R²).

Together, these analyses allowed us to identify which audio features were most strongly associated with variations in dancers' movement intensity and to assess the strength of the coupling between movement and audio dynamics produced by the system.

Results. Our analysis revealed several key relationships between movement energy and audio features generated by the system.

Correlation Analysis. Pearson correlation analysis showed that the minimum movement energy (*min_energy*) had the strongest and most consistent relationships with audio features. In particular, *mfcc_1* (first Mel-frequency cepstral coefficient) exhibited a significant negative correlation with min_energy ($r = -0.45$, $p < 0.001$), suggesting that clips with lower minimum movement energy were associated with audio segments characterized by smoother spectral shapes. Spectral contrast in the sixth frequency band (*spec_contrast_6*) and *mfcc_7* also showed significant positive correlations with min_energy.

Principal Component Analysis (PCA). PCA indicated that a small number of components captured substantial variance in both movement and audio features. In particular, variations in energy-based movement measures (average, minimum, maximum, and standard deviation) loaded heavily onto the first few principal components, while audio variance was dominated by MFCCs and spectral features.

Canonical Correlation Analysis (CCA). CCA revealed moderate canonical correlations between the combined movement energy statistics and audio feature sets. The first canonical component pair linked high standard deviation of movement energy with variations in spectral complexity and dissonance in the audio, indicating multivariate coupling between movement expressivity and audio texture.

Partial Least Squares (PLS) Regression. PLS regression models found that audio features could modestly predict movement energy, with the highest R² value (0.103) for predicting *min_energy*. Conversely, movement features showed stronger predictive power for certain audio characteristics: movement energy metrics could predict *mfcc_1* with an R² of 0.202 and *mfcc_7* with an R² of 0.162, suggesting that dancers' movement dynamics influenced the timbral qualities of the generated soundtrack.

Random Forest Feature Importance. Random forest regressions further confirmed the importance of specific audio features. *mfcc_1*, *mfcc_7*, *spec_contrast_6*, and *spectral_flux* consistently emerged as the most important predictors of movement energy statistics across models, aligning with the findings of the linear analyses.

Overall, the results suggest that the soundscape created by our system responded most consistently to variations in dancers' minimum movement energy, with audio features related to timbre and spectral dynamics (MFCCs and spectral contrast) showing the strongest associations with movement intensity.

Qualitative results Dancers reported a fluid exchange of initiative with the system throughout their performances. The system often influenced their movement choices at both macro and micro levels, inspiring exploration of new dance expressions to discover corresponding musical responses. Conversely, dancers sometimes deliberately attempted to redirect the musical atmosphere, such as introducing energetic movement during ambient passages. While the system typically required 5-10 seconds to adapt to significant changes in dance energy, this delay was intentionally designed to maintain musical coherence. This adaptation period could potentially be customized based on dance genre preferences, balancing responsiveness against musical continuity.

Conclusion

This research presents a novel approach to dance through AI-mediated co-creation, fundamentally reimagining the traditional relationship between movement and music. Our system enables dancers to dynamically shape musical environments while simultaneously responding to them, creating a bidirectional creative partnership where initiative flows fluidly between human and machine. Statistical analysis confirms meaningful correlations between specific movement qualities and audio features.

This work explores a multi-layered creative relationship. The original composer's musical intent—which may itself be improvisational—becomes raw material for a new emergent composition, dynamically rearranged and remixed through the dancer's movements. The resulting sonic experience is a form of real-time collage where three creative forces converge: the original compositional elements, the system's algorithmic decision-making, and the dancer's embodied expression. Each performance thus represents a unique relationship between these creative entities, with the dancer physically sculpting a new musical composition from fragments of the composer's work, creating something neither could have produced independently.

By inverting the traditional paradigm where movement follows sound, we position dancers as active co-creators of

[1] Detailed statistical analysis results are provided on the supplementary-material website: https://sites.google.com/view/reimagining-dance

the musical arrangements. The dance movements can serve as a novel compositional tool for creating musical content that has artistic value beyond the performance context. This approach offers new creative possibilities not only to dance performance, but also as a form of musical composition.

While the system was primarily envisioned as a co-creative artistic framework for dance performance, another promising application emerged during our pilot study. We observed that when participant dancers experienced fatigue and naturally reduced their movement intensity, the system organically transitioned from dynamic rhythmic music to more ambient soundscapes. This adaptive quality could be valuable not only in dance performance, but also in exercise and training settings.

We are currently planning a large-scale study with a professional dance company to evaluate the system's impact on choreographic process and audience reception. Through this ongoing research, we aim to develop a deeper understanding of the emergent creative language that evolves between human dancers and AI musical collaborators.

References

Bailey, D. 1992. *Improvisation: Its Nature and Practice in Music*. Da Capo Press.

Carter, C. L. 2000. Improvisation in dance. *Journal of Aesthetics and Art Criticism* 58(2):181–190.

Colton, S., and Wiggins, G. A. 2012. Computational creativity: The final frontier? In *ECAI 2012*, volume 242 of *Frontiers in Artificial Intelligence and Applications*, 21–26. IOS Press.

De Spain, K. 2014. *Landscape of the Now: A Topography of Movement Improvisation*. Oxford University Press.

Deterding, S.; Hook, J.; Fiebrink, R.; Gillies, M.; Gow, J.; Akten, M.; Smith, G.; Liapis, A.; and Compton, K. 2017. Mixed-initiative creative interfaces. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, CHI EA '17, 628–635. New York, NY, USA: ACM.

Foster, S. 2002. *Dances that Describe Themselves: The Improvised Choreography of Richard Bull*. Wesleyan University Press.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, 2672–2680.

Hoffman, G., and Weinberg, G. 2011. Interactive improvisation with a robotic marimba player. *Autonomous Robots* 31(2):133–153.

Jennings, K. E. 2010. Developing creativity: Artificial barriers in artificial intelligence. *Minds and Machines* 20(4):489–501.

Kingma, D. P., and Welling, M. 2014. Auto-encoding variational Bayes. In *Proceedings of the 2nd International Conference on Learning Representations (ICLR)*.

Kumar, M.; Long, D.; and Magerko, B. 2020. Creativity metrics for a lead-and-follow dynamic in an improvisational dance agent. In *Proceedings of the International Conference on Computational Creativity*.

Lee, H.-Y.; Yang, X.; Liu, M.-Y.; Wang, T.-C.; Lu, Y.-D.; Yang, M.-H.; and Kautz, J. 2019. Dancing to music. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, 3586–3596. Red Hook, NY, USA: Curran Associates Inc.

Nachmanovitch, S. 1990. *Free Play: Improvisation in Life and Art*. Penguin Putnam.

Tang, T.; Jia, J.; and Mao, H. 2018. Dance with melody: An lstm-autoencoder approach to music-oriented dance synthesis. In *Proceedings of ACM Multimedia Conference*.

Tsuchida, S.; Fukayama, S.; Hamasaki, M.; and Goto, M. 2019. AIST dance video database: Multi-genre, multi-dancer, and multi-camera database for dance information processing. In *Proceedings of the 20th International Society for Music Information Retrieval Conference, ISMIR 2019*.

Winship, L. 2024. Small step or a giant leap? what AI means for the dance world. *The Guardian*. Accessed: January 11, 2025.