DuetGen: Music Driven Two-Person Dance Generation via Hierarchical Masked Modeling

ANINDITA GHOSH, Saarland University
DFKI, Germany and Saarland University
Max Planck Institute for Informatics, Germany
BING ZHOU, Snap Inc., USA
RISHABH DABRAL, Max Planck Institute for Informatics, Germany
JIAN WANG, Snap Inc., USA
VLADISLAV GOLYANIK, Max Planck Institute for Informatics, Germany
CHRISTIAN THEOBALT, Max Planck Institute for Informatics, Germany
PHILIPP SLUSALLEK, DFKI, Germany and Saarland University, Germany
CHUAN GUO, Snap Inc., USA

Fig. 1. **DuetGen** generates synchronized two-person dance choreography from input music, featuring natural and close interactions between dancers.

We present DuetGen, a novel framework for generating interactive two-person dances from music. The key challenge of this task lies in the inherent complexities of two-person dance interactions, where the partners need to synchronize both with each other and with the music. Inspired by the recent advances in motion synthesis, we propose a two-stage solution: encoding two-person motions into discrete tokens and then generating these tokens from music. To effectively capture intricate interactions, we represent both dancers' motions as a unified whole to learn the necessary motion tokens, and adopt a *coarse-to-fine* learning strategy in both the stages. Our first stage utilizes a VQ-VAE that hierarchically separates high-level semantic features at a coarse temporal resolution from low-level details at a finer resolution, producing two discrete token sequences at different abstraction levels. Subsequently, in the second stage, two generative masked transformers learn to map music signals to these dance tokens: the first producing high-level semantic tokens, and the second, conditioned on music and these semantic tokens, producing the low-level tokens. We train both transformers to learn to predict randomly masked tokens within the sequence, enabling them to iteratively generate motion tokens by filling an empty token sequence during inference. Through the hierarchical masked modeling and dedicated interaction representation, DuetGen achieves the generation of synchronized and interactive two-person dances across various genres. Extensive experiments and user studies on a benchmark duet dance dataset demonstrate state-of-the-art performance of DuetGen in motion realism, music-dance alignment, and partner coordination. Code and model weights are available at https://github.com/anindita127/DuetGen.

CCS Concepts: • **Computing methodologies** → **Procedural animation**.

Additional Key Words and Phrases: Motion Synthesis, Music to Dance Synthesis, Two-Person Motion, Couple Dance, Duet Dance, Motion Tokenization, Masking, Transformer

ACM Reference Format:
Anindita Ghosh, Bing Zhou, Rishabh Dabral, Jian Wang, Vladislav Golyanik, Christian Theobalt, Philipp Slusallek, and Chuan Guo. 2025. DuetGen: Music

*Work done during internship at Snap Inc.
Authors' Contact Information: Anindita Ghosh, Saarland University
DFKI, Saarbrücken, Saarland, Germany and Saarland University
Max Planck Institute for Informatics, Saarbrücken, Saarland, Germany, anindita.ghosh@dfki.de; Bing Zhou, Snap Inc., New York City, NY, USA, bzhou@snapchat.com; Rishabh Dabral, Max Planck Institute for Informatics, Saarbrücken, Germany, rdabral@mpi-inf.mpg.de; Jian Wang, Snap Inc., New York City, NY, USA, jwang4@snapchat.com; Vladislav Golyanik, Max Planck Institute for Informatics, Saarbrücken, Germany, golyanik@mpi-inf.mpg.de; Christian Theobalt, Max Planck Institute for Informatics, Saarbrücken, Germany, theobalt@mpi-inf.mpg.de; Philipp Slusallek, DFKI, Saarbrücken, Germany and Saarland University, Saarbrücken, Germany, philipp.slusallek@dfki.de; Chuan Guo, Snap Inc., New York City, NY, USA, cguo2@snapchat.com.

SIGGRAPH Conference Papers '25, Vancouver, BC, Canada

ACM ISBN 979-8-4007-1540-2/25/08
https://doi.org/10.1145/3721238.3730741

Driven Two-Person Dance Generation via Hierarchical Masked Modeling. In *Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers (SIGGRAPH Conference Papers '25), August 10–14, 2025, Vancouver, BC, Canada.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3721238.3730741

1 Introduction

Duet dancing is a fundamental aspect of human culture that embodies coordination and mutual expression between partners. As one of the most interactive and visually engaging dance forms [Raheb et al. 2019], it holds a significant presence across traditional and contemporary art. With the proliferation of contemporary art in digital platforms today, the capability to synthesize dances, including duet dances, fosters new opportunities for creative content generation in domains spanning entertainment and communication media, virtual social interactions, and virtual education, among others. While recent advances in text-based motion generation [Ghosh et al. 2021; Guo et al. 2024a; Petrovich et al. 2022] and editing [Athanasiou et al. 2024; Li et al. 2025] allow users to choreograph dance through natural language, music offers a more natural and temporally aligned modality for generating expressive and synchronized duet performances [Gong et al. 2023].

Unlike single-person dances [Alexanderson et al. 2023; Bhattacharya et al. 2024; Tseng et al. 2023], where timing and spatial consistency concern only one body, two-person dance synthesis gives rise to several key challenges in motion modeling. It requires maintaining synchronized motions between the dancers to ensure cohesive rhythm, tempo, and mutual positioning (as illustrated in Figure 1). It further demands intricate modeling of lead-follow dynamics and close interactions, including twirls, lifts, and partner supports across diverse dance styles — all while aligning both dancers' motions with the music. Recent works have explored group-dance synthesis [Le et al. 2023a,b; Yao et al. 2023], focusing on synchronized multi-person dance motions. However, these methods do not address close interactions essential in duet dancing. While Duolando [Siyao et al. 2024] pioneered dance generation in two-person scenarios, it is limited to reactive partner motions rather than simultaneously generating motions of both dancers.

Towards this goal, we introduce *DuetGen*, the *first* method designed to generate synchronized and interactive two-person dance motions from music. Our key insights are twofold. First, we propose a unified representation for two-person motions that treats both dancers' movements as a whole, rather than modeling individual motions separately using single-person representations [Javed et al. 2024; Siyao et al. 2024]. This unified representation proves especially effective in reducing the modeling complexity of inter-person interactions. Second, we propose a hierarchical learning of the unified dance motions, encoding movements into two levels of discrete motion tokens representing global semantics and fine details, respectively. This allows for efficient learning of the individual and interactive motions, and synchronizing them with music. To implement our approach, we train a hierarchical VQ-VAE model to learn motion tokens at the two levels. We then employ two transformers to generate the motion tokens from music. The first transformer generates high-level semantic tokens from music, and the second transformer, conditioned on both music and the high-level tokens, generates the low-level tokens. We train both transformers using a masked modeling approach such that during inference, the transformers can iteratively predict a complete sequence of motion tokens from a fully-masked masked sequence. We also incorporate a root motion predictor to refine the global root trajectories in the generated motions.

In summary, our main contributions are the following.

- DuetGen, the first framework for generating interactive two-person dance motions directly from music.
- A technique to model two-person dance motions in the discrete space with a hierarchy of abstractions, using our unified two-person representation, multi-scale motion quantization and two-stage generative masked transformers.
- Extensive experiments and user studies on the benchmark DD100 dataset [Siyao et al. 2024], demonstrating our effectiveness in synthesizing realistic two-person dances.

2 Related Work

We briefly summarize prior work in the related areas of motion quantization, multi-person motions, and music-to-dance synthesis.

Deep Motion Quantization. Aristidou et al. [Aristidou et al. 2018] utilize triplet contrastive learning to convert continuous motions into semantically meaningful discrete motif words. Vector-quantized VAE (VQ-VAE) [Van Den Oord et al. 2017] was first applied to human motion modeling in TM2T [Guo et al. 2022b], where autoencoded motion latent codes were quantized to indexed entries in a learnable codebook. T2M-GPT [Zhang et al. 2023] enhanced this approach through exponential moving average and code reset techniques, which was then adopted by subsequent works [Jiang et al. 2023; Zhang et al. 2024a]. SynNsync [Maluleke et al. 2024] quantizes full-body motion by disentangling motion into its constituent pose, orientation, and translation via a parametric body model. To reduce quantization error, MoMask [Guo et al. 2024a] implemented residual quantization (RVQ) [Borsos et al. 2023] and used multiple layers to iteratively quantize a vector and its residuals. While RVQ achieves improved motion-space mapping, its quantization layers operate at full temporal scale and the tokens describe quantization residuals that have no specific motion semantics. In contrast, we follow a coarse-to-fine strategy that naturally separates high-level semantic features from low-level details, and represent them as motion tokens at different temporal resolutions.

Multi-Person Motion Synthesis. Synthesizing motions for multiple persons presents unique challenges in modeling their interactions [Sui et al. 2025]. Starke et al. [Starke et al. 2020, 2021] pioneered the use of mixture-of-expert networks to autoregressively predict poses based on inter-character features and instant control signals. Guo et al. [Guo et al. 2022a] introduced transformer-based cross-attention networks to forecast intensive interactions, such as combat, from historical observations. A separate body of research has emerged on reactive motion synthesis [Chopin et al. 2023; Ghosh et al. 2024; Liu et al. 2023; Siyao et al. 2024; Xu et al. 2024b], where one character's movements are generated in response to others. Notable contributions include InterFormer's skeleton-aware spatial attention [Chopin et al. 2023] and the integration of human-object

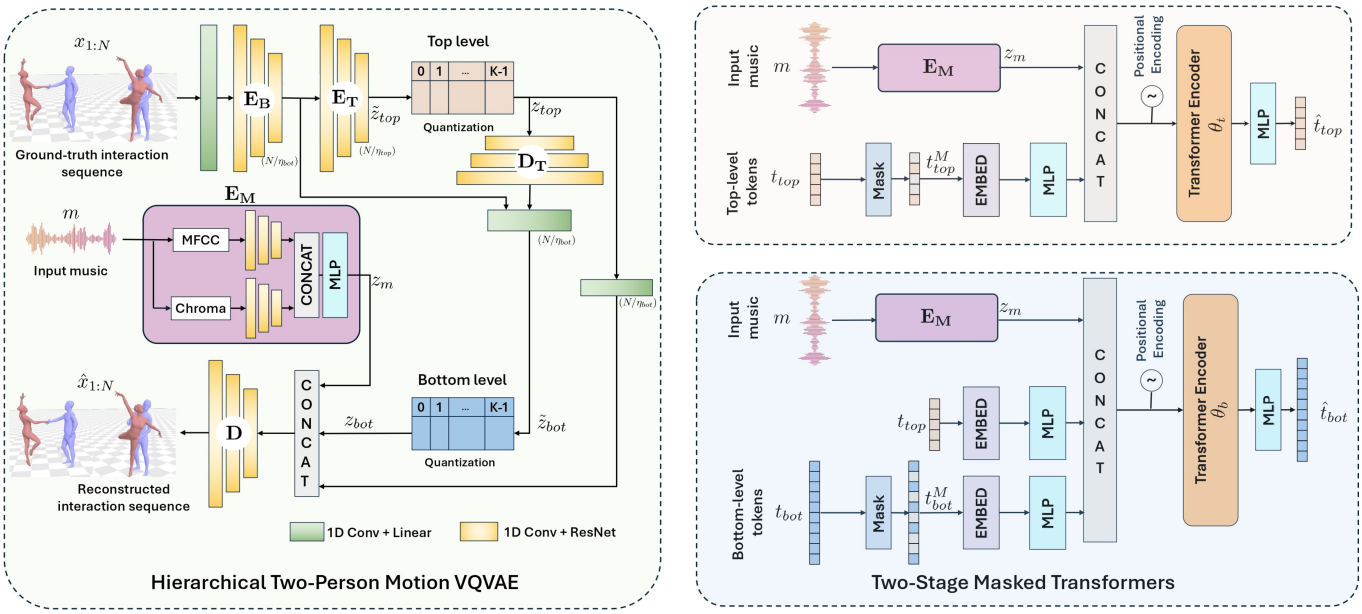

Fig. 2. **DuetGen Training Framework.** *Left:* Our hierarchical two-person motion VQ-VAE encodes a unified two-person motion sequence x of length N into two-scale discrete token sequences. Top-level tokens at a coarse temporal resolution (N/η_{top}) capture global semantics, and bottom-level tokens at finer temporal resolution (N/η_{bot}) capture complementary low-level details. *Right:* We employ two transformers to model tokens of different granularity via masked modeling. The first-stage transformer (*top-right*) predicts the masked top-level tokens from music. Subsequently, the second-stage transformer (*bottom-right*) predicts the masked bottom-level tokens conditioned on both the music and the top-level token sequence.

interactions [Liu et al. 2023]. More recently, ReMoS [Ghosh et al. 2024] and ReGenNet [Xu et al. 2024b] have improved the generation of interpersonal dynamics using combined diffusion-transformer architectures and interaction losses. Text-driven multi-person motion generation has also gained traction, evolving from approaches that fine-tune single-person diffusion priors [Shafir et al. 2023; Wang et al. 2024] to more sophisticated methods [Liang et al. 2024; Ponce et al. 2024] enabled by large datasets like InterHuman [Liang et al. 2024] and Inter-X [Xu et al. 2024a]. Recent works [Fan et al. 2024; Shan et al. 2024] have extended these capabilities to more than two persons through specialized transformer architectures and pose-based intermediaries. An emerging alternative leverages discrete representations, typically in a two-stage process: motion-to-token conversion [Van Den Oord et al. 2017] followed by token generation. This approach has shown promise in applications such as Duolando's audio-guided partner dance generation [Siyao et al. 2024] and InterMask's text-driven two-person motion synthesis [Javed et al. 2024]. We further extend this approach to duet dances using a unified representation of two-person motions and a hierarchical tokenization.

Music-Driven Dance Synthesis. Traditional music-to-dance synthesis approaches [Arikan and Forsyth 2002; Au et al. 2022; Chen et al. 2021; Shiratori et al. 2006] relied on motion graphs and dynamic time warping to retrieve matching motion clips from existing libraries. While these methods generate plausible movements, they are constrained by their reliance on pre-existing motions and the need to maintain extensive motion libraries, limiting their scalability. The advent of deep learning shifted the paradigm toward data-driven approaches, framing music-to-dance as an autoregressive problem and solving it using RNNs and transformers [Huang et al. 2020; Li et al. 2020, 2021; Sun et al. 2022]. While these models effectively learned movement patterns and music correspondences, their deterministic nature often led to repetitive or frozen motions. Subsequent explorations on deep generative models, particularly with GAN-based approaches [Li et al. 2022; Sun et al. 2020] and hybrid GAN-VAE models [Lee et al. 2019] for composing long-term dances from shorter clips, alleviated some of these limitations. A significant breakthrough came with diffusion models [Ho et al. 2020] and GPT models [Brown et al. 2020], which substantially improved motion quality and realism [Alexanderson et al. 2023; Siyao et al. 2022; Tseng et al. 2023]. This success led to specialized architectures like directional autoregressive diffusion [Zhang et al. 2024b], coarse-to-fine generation [Li et al. 2024a], and full-body synthesis [Li et al. 2023]. Bailando [Siyao et al. 2022] mapped dance movements to discrete tokens using VQ-VAE and generated them autoregressively with GPT. This discrete representation framework was extended by TM2D [Gong et al. 2023] to achieve music-to-dance synthesis with further controllability from text prompts by sharing the codebooks. MoFusion [Dabral et al. 2023], and UDE [Zhou and Wang 2023] also approached this bimodality driven 3D dance generation, but with diffusion models. Nevertheless, these works remain limited to single-person dance synthesis. While several works have explored generating group choreography from music [Le et al. 2023a,b; Wang et al. 2022], they typically neglect close interactions between dancers. Duolando [Siyao et al. 2024] and InterDance [Li et al. 2024b] made initial progress by generating follower movements in response to a

given leader's sequence, with InterDance having the capability to generate two-person motion from audio by replacing the leader's motion with random noise. Our work advances this field by simultaneously generating coordinated movements and close interactions for both dancers attuned to music signals.

3 Method

To generate two-person dance motions from music signal, we first model the dance motions in a discrete space. Figure 2 illustrates the training pipeline of our method, which consists of two main components: a hierarchical motion VQ-VAE (Section 3.2) that transforms motions into discrete tokens, and two-stage masked transformers (Section 3.3) that maps input music to these tokens. After training, the full music-to-dance mapping is achieved through the inference procedure shown in Figure 3.

3.1 Data Representation

Unified Two-Person Motion Representation. Without loss of generality, we denote the two partners in a duet dance as $\mathcal{A}$ and $\mathcal{B}$ (note that the designations $\mathcal{A}$ and $\mathcal{B}$ are interchangeable). We represent the motions of a duet dance as $x_{1:N} \in \mathbb{R}^{N\times D}$ with N frames and D dimensional motion features that combine the motions of $\mathcal{A}$ and $\mathcal{B}$. We adopt the SMPL [Loper et al. 2023] body structure with $J = 22$ body joints. The unified two-person interaction representation for the i^{th} motion frame consists of

$$x_i = \left[t^{\delta}_{\mathcal{A}}, r^{g}_{\mathcal{A}}, j^{p}_{\mathcal{A}}, j^{r}_{\mathcal{A}}, j^{v}_{\mathcal{A}}, c^{f}_{\mathcal{A}}, t^{\epsilon}_{\mathcal{B}}, r^{g}_{\mathcal{B}}, j^{p}_{\mathcal{B}}, j^{r}_{\mathcal{B}}, j^{v}_{\mathcal{B}}, c^{f}_{\mathcal{B}} \right]_i, \quad (1)$$

where $t^{\delta}_{\mathcal{A}} \in \mathbb{R}^3$ is the root translation difference for $\mathcal{A}$ from *their own previous frame*, $t^{\epsilon}_{\mathcal{B}} \in \mathbb{R}^3$ is the root translation difference for $\mathcal{B}$ *relative to* $\mathcal{A}$*'s current frame*, $r^{g}_{\mathcal{A}}, r^{g}_{\mathcal{B}} \in \mathbb{R}^6$ are the global orientations of the root joint $\mathcal{A}$ and $\mathcal{B}$, $j^{p}_{\mathcal{A}}, j^{p}_{\mathcal{B}} \in \mathbb{R}^{(J-1)\times 3}$ are their root-invariant local joint positions, $j^{r}_{\mathcal{A}}, j^{r}_{\mathcal{B}} \in \mathbb{R}^{(J-1)\times 6}$ are their 6D local joint rotations, $j^{v}_{\mathcal{A}}, j^{v}_{\mathcal{B}} \in \mathbb{R}^{J\times 3}$ are their local joint velocities, and $c^{f}_{\mathcal{A}}, c^{f}_{\mathcal{B}} \in \mathbb{R}^4$ are their binary foot-ground contact indicators obtained by thresholding the heel and toe joint velocities. Together, these features add up to the motion feature dimension $D = 536$.

Note that we deliberately represent $\mathcal{A}$'s global position in the global coordinates as velocity and represent $\mathcal{B}$'s global position relative to $\mathcal{A}$'s. This *relational global positioning* ensures that $\mathcal{B}$'s global position is informed by their partner, making its prediction more manageable for networks. Our experiments demonstrate the key role of this approach in both VQ learning (Table 1) and the final generation (Table 2).

Music Features. We sample the input music m at a rate that matches the temporal resolution of the motion sequence. We then extract MFCC [Hossan et al. 2010], MFCC delta, and Chroma features [Shah et al. 2019] using the *Librosa* library. The MFCC features are 40-dimensional and the Chroma features are 12-dimensional.

3.2 Hierarchical Two-Person Motion Quantization

As illustrated in Figure 2 (left), we discretize two-person dance motions into a hierarchy of tokens. The top-level tokens model high-level motion semantics, *e.g.*, actions such as walking and turning induced by overall body movements, and bottom-level tokens represent fine-grained details, *e.g.*, intricate articulations specific to a dance choreography. To obtain these tokens, our VQ-VAE encoder network consists of a bottom encoder $\mathbf{E_B}$ followed by a top encoder $\mathbf{E_T}$. Both $\mathbf{E_B}$ and $\mathbf{E_T}$ consist of 1D convolution blocks and they temporally downscale the input motion $x_{1:N}$ by factors of η_{bot} and η_{top} respectively, where $\eta_{top} > \eta_{bot} > 1$. $\mathbf{E_T}$ yields latent code sequences $\tilde{z}_{top} \in \mathbb{R}^{(N/\eta_{top})\times D_{top}}$, where D_{top} denotes the spatial dimensionality. We quantize this code sequence by replacing each of the code with its nearest entry in a learnable codebook $C_{top} = \left\{c^{t}_{k}\right\}_{k=1}^{K} \subset \mathbb{R}^{D_{top}}$ consisting of K codes, resulting in the quantized vector sequence $z_{top} \in \mathbb{R}^{(N/\eta_{top})\times D_{top}}$. Specifically,

$$z_{top} = \mathbf{Q_T}(\tilde{z}_{top}); \quad \tilde{z}_{top} = \mathbf{E_T}(\mathbf{E_B}(x_{1:N})), \quad (2)$$

where $\mathbf{Q_T}(\cdot)$ denotes the quantization operation at the top level.

Subsequently, we upscale the top-level quantized features $z_{top} \in \mathbb{R}^{(N/\eta_{top})\times D_{top}}$ by a factor of η_{top}/η_{bot} through the top decoder $\mathbf{D_T}$. We combine them with the output features from $\mathbf{E_B}$ to produce latent features at a finer temporal resolution $\tilde{z}_{bot} \in \mathbb{R}^{(N/\eta_{bot})\times D_{bot}}$, where D_{bot} denotes the spatial dimensionality. We quantize these bottom-level latent codes using a separate codebook $C_{bot} = \left\{c^{b}_{k}\right\}_{k=1}^{K} \subset \mathbb{R}^{D_{bot}}$ where K is the number of codes in the codebook, yielding quantized feature sequences $z_{bot} \in \mathbb{R}^{(N/\eta_{bot})\times D_{bot}}$, as

$$\tilde{z}_{bot} = \left(\mathbf{E_B}\left(x_{1:N}\right), \mathbf{D_T}\left(z_{top}\right)\right); \quad z_{bot} = \mathbf{Q_B}(\tilde{z}_{bot}), \quad (3)$$

where $\mathbf{Q_B}(\cdot)$ indicates the bottom-level quantization process.

We then concatenate the quantized features z_{top} and z_{bot}, and decode them back to the motion sequence $\hat{x}_{1:N} \in \mathbb{R}^{N\times D}$ via the decoder $\mathbf{D}$. Interestingly, we observe that conditioning the decoder $\mathbf{D}$ additionally on the music features z_m encourages faster convergence of the training and better VQ-VAE reconstructions (Table 1). Overall, the final reconstructed two-person sequence is given as

$$\hat{x}_{1:N} = \mathbf{D}\left(z_{top}, z_{bot}, z_m\right). \quad (4)$$

Training Objectives. We train the hierarchical VQ-VAE framework end-to-end. Our primary VQ-VAE training objective consists of a motion reconstruction loss $\mathcal{L}_r = \|x - \hat{x}\|_2$ and a motion velocity loss $\mathcal{L}_v = \|\Delta x - \Delta\hat{x}\|_2$. Following T2M-GPT[Zhang et al. 2023], we use the exponential moving average (EMA) and codebook reset (Code Reset) techniques for both C_{top} and C_{bot} codebooks. We optimize for commitment losses on the two quantized latent sequences, as

$$\mathcal{L}_{com} = \beta_1 \left\|\tilde{z}_{top} - sg\left[z_{top}\right]\right\|_2 + \beta_2 \left\|\tilde{z}_{bot} - sg\left[z_{bot}\right]\right\|_2, \quad (5)$$

where β_1 and β_2 are hyper-parameters for the commitment loss and $sg[\cdot]$ is the stop-gradient operator.

While optimizing for overall character movements, the aforementioned losses do not enforce synchronization fidelity between the two dancers. Therefore, we apply feature transformation and forward kinematics $fk(\cdot)$ to the ground-truth motions features x and the reconstructed motion features $\hat{x}$ to obtain the joint positions of $\mathcal{A}$ and $\mathcal{B}$ in the global coordinates, as $(p_{\mathcal{A}}, p_{\mathcal{B}}) = fk(x)$ and $(\hat{p}_{\mathcal{A}}, \hat{p}_{\mathcal{B}}) = fk(\hat{x})$ respectively. We then enforce reconstruction loss

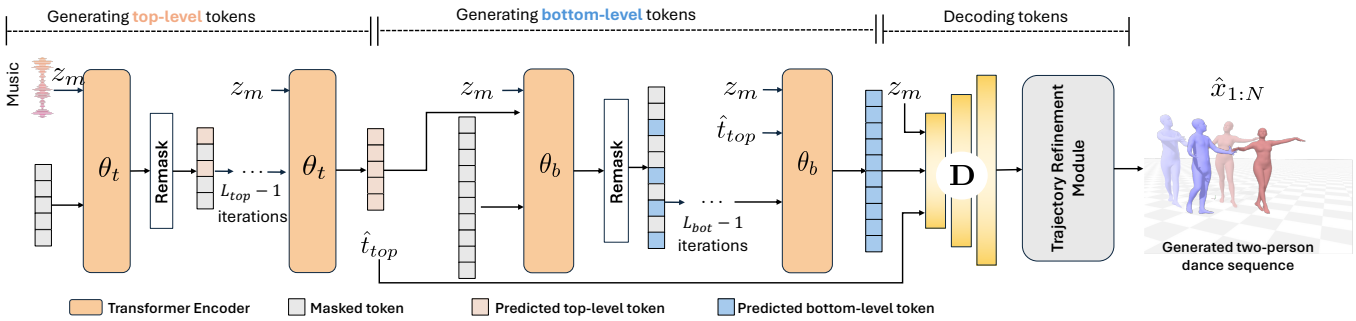

Fig. 3. **Inference Process.** Our first-stage transformer θ_t iteratively fills an empty sequence of top-level tokens in L_{top} iterations based on input music. Then, the second-stage transformer θ_b generates the complete sequence of bottom-level tokens in L_{bot} iterations, conditioned on both music and the generated top-level tokens $\hat{t}_{top}$. We combine these token sequences and decode them into two-person dance motions via the VQ-VAE decoder D. We then apply a lightweight network to refine the root trajectories and produce the final outputs. For brevity, we only show the main network modules here.

on the global-coordinate joint positions and relative joint distances between the two-persons, as

$$\mathcal{L}_{fk} = \|p_{\mathcal{A}} - \hat{p}_{\mathcal{A}}\|_2 + \|p_{\mathcal{B}} - \hat{p}_{\mathcal{B}}\|_2, \tag{6}$$

$$\mathcal{L}_{rel} = \frac{1}{J}\sum_{j\in J}\lambda_j \sum_{k\in J} e^{-d\left(p_{\mathcal{A}_j}, p_{\mathcal{B}_k}\right)} \left| d\left(p_{\mathcal{A}_j}, p_{\mathcal{B}_k}\right) - d\left(\hat{p}_{\mathcal{A}_j}, \hat{p}_{\mathcal{B}_k}\right)\right|, \tag{7}$$

where $d(\cdot,\cdot)$ is the Euclidean distance computed between each joint positions of $\mathcal{A}$ and $\mathcal{B}$, and λ_j is a variable weight applied to individual joints with higher values for the end-effectors to better learn their rapid movements. We also use an exponentially decaying distance-aware weight $e^{-d\left(p_{\mathcal{A}_j}, p_{\mathcal{B}_k}\right)}$ at each frame to emphasize joints that are closer to the other person and therefore more relevant for interactions [Ghosh et al. 2024].

Overall, we train our hierarchical two-person VQ-VAE to minimize the weighted sum of all loss terms, as

$$\mathcal{L}_{vq} = \lambda_r\mathcal{L}_r + \lambda_v\mathcal{L}_v + \lambda_{com}\mathcal{L}_{com} + \lambda_{fk}\mathcal{L}_{fk} + \lambda_{rel}\mathcal{L}_{rel}, \tag{8}$$

where λ_r, λ_v, λ_{com}, λ_{fk} and λ_{rel} are scalar weights to balance the individual loss components.

Discrete Representation. After training, we can represent each code in the quantized sequence at the top level, z_{top}, by its corresponding index k in the codebook C_{top}, thus transforming a two-person motion sequence into a discrete token sequence $t_{top} \in \{0, 1, ..., K\}^{N/\eta_{top}}$. Similarly, we obtain the bottom-level tokens as $t_{bot} \in \{0, 1, ..., K\}^{N/\eta_{bot}}$. We also retain the ability to decode these token sequences back to two-person motions. Our following two-stage transformers learn to generate this hierarchy of discrete tokens from music through generative masked modeling.

3.3 Music to Motion Via Two-Stage Masked Transformers

We use two-stage bidirectional transformers to map input music signals m to the discrete two-person dance token sequences t_{top} and t_{bot} (Figure 2, right). To this end, we follow a generative masked modeling approach. First, we randomly mask out a varying fraction of sequence elements by replacing the original tokens with a special [MASK] token. We then train the transformers to predict these masked tokens given the token context and relevant conditions. Following MaskGIT [Chang et al. 2022], we adopt a cosine function for scheduling the masking ratio as $\gamma(\tau) = \cos\left(\frac{\pi\tau}{2}\right) \in [0, 1]$ where $\tau \sim \mathcal{U}(0, 1)$. During training, we randomly sample τ at each iteration, and uniformly select $\lceil\gamma(\tau)\cdot N\rceil$ tokens from the token sequence of length N for masking. To further enhance the contextual reasoning of the masked transformer, we adopt the re-masking strategy [Guo et al. 2024a] used in BERT pre-training [Devlin et al. 2018], where we either replace a small portion of masked tokens with random tokens or restore them to their original values.

Modeling Top-Level Motion Tokens. We first encode the music signal through the music encoder $\mathbf{E_M}$ to obtain music features z_m. Following the masking strategy, we mask out portions of the top-level tokens to get t^M_{top}, which we embed and then concatenate with z_m. After positional encoding, we provide these features into a transformer encoder θ_t, which predicts the discrete distribution of tokens at the masked positions. Our objective is to minimize the negative log-likelihood of the prediction on masked tokens, as

$$\mathcal{L}_{tmask} = \sum_{t^M_{top_k}=[\text{MASK}]} -\log\theta_t\left(t_{top_k}\middle| t^M_{top}, m\right). \tag{9}$$

Modeling Bottom-Level Motion Tokens. The bottom-level tokens are modeled by a separate transformer θ_b in a similar fashion, with the addition of conditioned on the associated top-level tokens. Following the same masking strategy, we mask portions of the bottom-level tokens to get t^M_{bot}. The training objective is to minimize the negative log-likelihood of the predicted masked tokens conditioned on the bottom-level token contexts t^M_{bot}, full sequence of top-level tokens t_{top}, and music signals m:

$$\mathcal{L}_{bmask} = \sum_{t^M_{bot\,k}=[\text{MASK}]} -\log\theta_b\left(t_{bot\,k}\middle| t^M_{bot}, m, t_{top}\right). \tag{10}$$

For both transformers, we adopt a classifier-free guidance strategy (CFG) [Chang et al. 2022], where we train the transformers without music feature conditioning for 10% of the time.

3.4 Inference

During inference, we generate a two-person dance sequence $\hat{x}$ of length N from input music (Figure 3). We start from an empty sequence of length (N/η_{top}), where all the tokens are masked, and use θ_t conditioned on music feature z_m to generate the top-level token sequence in L_{top} iterations. At each iteration l, θ_t predicts the probability distribution of tokens at the masked locations and samples motion tokens accordingly. We re-mask the sampled tokens with the lowest $\left\lceil \gamma\left(\frac{l}{L_{top}}\right) \cdot \left(\frac{N}{\eta_{top}}\right) \right\rceil$ probabilities, while keeping the other tokens unchanged for subsequent iterations. This process repeats until iteration L_{top}, yielding the complete sequence of top-level tokens $\hat{t}_{top}$. Next, we use θ_b to generate the bottom-level token sequence $\hat{t}_{bot}$ of length (N/η_{bot}) through L_{bot} iterations similarly, conditioned on both z_m and $\hat{t}_{top}$. Throughout this process, we apply classifier-free guidance at the projection layer of the transformers with guidance scale s. Finally, we decode these two motion token sequences to the motion space using the VQ-VAE decoder $\mathbf{D}$.

Trajectory Refinement Module. While our method generates plausible dance motions, the results exhibit noticeable sliding issues, primarily due to inaccurate root motions. Since global root motions can generally be determined from local body movements [Guo et al. 2024b], we implement a regressor with 1D convolutional layers. This regressor takes root orientations and local motion features $x_i^{local} = \left[r^g_{\mathcal{A}}, j^p_{\mathcal{A}}, j^r_{\mathcal{A}}, j^v_{\mathcal{A}}, c^f_{\mathcal{A}}, r^g_{\mathcal{B}}, j^p_{\mathcal{B}}, j^r_{\mathcal{B}}, j^v_{\mathcal{B}}, c^f_{\mathcal{B}}\right]_i$ as input and predicts root trajectory velocities $x_i^{traj} = \left[t^{\delta}_{\mathcal{A}}, t^{\epsilon}_{\mathcal{B}}\right]_i$. We train the regressor using reconstruction loss on both absolute root positions and root velocities. After training, we apply this network to the generated motions and replace the original root trajectories with the predictions based on the generated local motions.

4 Experiments

We elaborate on our experimental setup, including dataset, baselines, ablations, and evaluation metrics. We provide the implementation details of our network components in the appendix.

4.1 Dataset and Baselines

We train and evaluate our model on the DD100 dataset [Siyao et al. 2024]. It is currently the most comprehensive dataset of 3D two-person dance, comprising 10 distinct dance genres featuring intricate interactions between the dancers and approximately 1.9 hours of two-person dance motion data in the SMPLX representation [Pavlakos et al. 2019] with corresponding music. The train-test split is 168,176 frames and 42,496 frames for two-person motions.

Data Pre-Processing. As a pre-processing step, we augment the original dataset by generating sub-sequences using a sliding window of 400 frames with a stride of 100 frames. We sample the accompanying music at 15,360 Hz to match the temporal dimension of the original motion sequence. We further augment the dataset by interchanging persons $\mathcal{A}$ and $\mathcal{B}$ to create mirrored samples. This results in a grand total of 4,556 training samples and 1,144 test samples, with a frame rate of 30 FPS. Finally, we process each two-person motion sequence by transforming the positions and orientations of $\mathcal{A}$ and $\mathcal{B}$ such that $\mathcal{A}$'s root joint is at the global origin in the first frame and $\mathcal{A}$'s body faces the Z_+ direction. We Z-normalize all the motion features before feeding them to the networks.

Baseline Setup. We select the most relevant motion synthesis methods adaptable to two-person dance generation: Duolando [Siyao et al. 2024], GCD [Le et al. 2023a], InterGen [Liang et al. 2024], and MoFusion [Dabral et al. 2023]. GCD was originally trained on the AIOZ-GDANCE dataset [Le et al. 2023b] for variable-size group choreography. We re-train it on DD100 for two-person dance generation with a thorough hyperparameter search and report the best performances. Duolando was initially designed to generate follower dance motions based on music signals and leader dance motions. To adapt it to two-person dance generation, we train a separate transformer model that generates the leader's dance sequence from the input music. We then use these generated leader's motions to condition the follower's motions using the Duolando framework with pre-trained weights on DD100 from the original work. InterGen, a diffusion-based approach, originally generates two-person motions from text prompts. To adapt its cooperative denoisers and mutual attention mechanisms for music-driven dance generation, we retrain it on DD100, replacing the original CLIP encoding [Radford et al. 2021] with our music encoder network. For MoFusion, we represent the two-person motion by concatenating the root-relative 3D joint position of each person. The second person is represented as a copy of the first person. The model is retrained on the DD100 dataset. Details on the preparation and training of these baselines are provided in the appendix.

4.2 Ablated Versions

We compare our proposed DuetGen with the following ablations.

- **A1: w/o relational global positioning.** Represents root positions of both persons' motions in the global coordinates, following conventional approaches [Liang et al. 2024].
- **A2: w/o hierarchical tokenization.** Tokenizes two-person dance motions using conventional single-layer VQ-VAE [Zhang et al. 2023] and generates tokens through a single masked transformer.
- **A3: w/o unified representation.** Tokenizes each person's motion separately [Javed et al. 2024], resulting in separate token sequences per person.
- **A4: w/o trajectory refinement module.** Removes the trajectory refinement module from our framework.
- **A5: alternate 438-D music encoding.** Uses 438-dimensional music representation [Siyao et al. 2024, 2022] instead of MFCC and Chroma features.
- **A6: w/o music feature z_m in VQ.** Trains the two-person VQ-VAE without conditional music signal encoding.
- **A7: alternate residual VQ.** Tokenizes the two-person dance sequences using four-layer Residual Quantization [Guo et al. 2024a], with a dropout ratio of 0.2.
- **A8: three-level token hierarchy.** Implements three hierarchies in the VQ-VAE.

4.3 Evaluation Metrics

We evaluate the two-person tokenizer's reconstruction quality using mean per joint position error (MPJPE) and mean per frame per joint velocity error (MPJVE) [Ghosh et al. 2024] on the reconstructed

motions after token decoding, along with the mean relative distance error (RDE) between the absolute root positions of $\mathcal{A}$ and $\mathcal{B}$. On dance generation, we evaluate models across three aspects: individual motion quality, synchronization between dancers, and music-dance alignment. For *individual motion quality*, we compute the averaged Fréchet Inception Distance (FID) [Heusel et al. 2017] over two persons, which measures the distributional difference between generated and ground-truth dance movement features. We also calculate latent variance to assess the diversity (Div) of generated dance movements, and the physical foot contact score (PFC) [Tseng et al. 2023] to measure the physical plausibility of the foot movements w.r.t. the ground plane. For *synchronization between dancers*, we use the PFID score [Ng et al. 2022], which compares the distributional difference between generated and ground-truth paired-motion features, instead of individual motion features. We calculate contact frequency percentage (CF), defined as the proportion of frames where dancers are in contact with each other. Contact is determined by a minimum absolute distance of less than 40 cm between any joints of the two dancers. For *music-dance alignment*, we use the Beat Alignment Score (BAS) following prior works [Bhattacharya et al. 2024; Siyao et al. 2024, 2022], and report the averaged score among two persons.

5 Results and Discussions

We summarize the results of our experiments and key observations.

5.1 Quantitative Evaluation

We evaluate our method quantitatively under two experimental settings: VQ reconstruction and two-person dance generation.

VQ Reconstruction. Table 1 reports the reconstruction quality of our hierarchical two-person motion VQ-VAE against the ablations and alternative designs discussed in Section 4.2. The vanilla VQ-VAE (ablation A2) produces noisy individual motion reconstructions (65.09 mm MPJPE) and struggles with high-fidelity close interactions (7.12 mm RDE). While applying 4-layer residual quantization in the middle of VQ-VAE (A7) partially mitigates these issues, the reconstruction quality remains suboptimal. In contrast, our hierarchical motion VQ-VAE, using only two quantization layers, demonstrates superior performance with high-fidelity reconstruction of both individual motions (36.26 mm MPJPE for person B) and interpersonal interactions (0.17 mm RDE). Our motion representation designs contribute significantly to this improvement. Specifically, incorporating *global relational positioning* and tokenizing *unified two-person motions* reduces individual motion per-joint reconstruction errors by over 30% and decreases the relative distance error from 7.12 to 0.17 millimeters. Further, our results demonstrate that incorporating conditional information (*e.g.*, music) in the VQ-VAE enhances its reconstruction performance. While we explored adding more hierarchical levels, introducing a third level of tokens (A8) yielded only marginal gains, with some metrics showing slight degradation, likely due to increased learning complexity.

Two-Person Dance Generation. Table 2 reports the performances of our method, the baselines, and the ablated versions on the DD100 dataset. The baselines are retrained on the DD100 dataset using publicly available codes for these models. DuetGen demonstrates superior performance across key metrics, achieving the lowest individual FID score (1.31 versus GCD's 9.71) and paired-FID score (2.54 versus GCD's 8.11). Our method also achieves the highest music-dance alignment with a BAS score of 0.215, while maintaining reasonable contact frequency on par with ground-truth dances. Ablation studies confirm the importance of both our hierarchical modeling (vs. ablation A2) and synchronization-dedicated representations (vs. A1 and A3). Removing either component leads to significant performance degradation, particularly in motion realism as measured by FID and PFID metrics. The two-hierarchy design provides the optimal balance between representational capacity and learning stability for duet dance generation, as opposed to training three-staged masked transformer (A8), where generation performance declines across key metrics. Finally, trajectory refinement (vs. A4) provides modest overall improvements, notably enhancing foot contact plausibility.

Table 1. **Quantitative Evaluation of VQ Reconstruction**. Comparison of motion reconstruction quality after tokenization across ablated versions. **Bold** indicates the best performance.

Method	MPJPE (mm) ↓		MPJVE (mm) ↓		RDE (mm) ↓
	Person $\mathcal{A}$	Person $\mathcal{B}$	Person $\mathcal{A}$	Person $\mathcal{B}$	
A1: w/o relational positioning	59.35	60.69	14.33	15.84	7.23
A2: w/o hier tokenization	58.35	65.09	15.38	17.54	7.12
A3: w/o unified representation	58.99	59.33	13.23	13.34	6.33
A6: w/o music feature z_m	39.60	43.71	12.82	14.68	1.32
A7: Residual VQ (4 layers)	49.77	57.62	12.51	14.49	4.54
A8: 3 level of hierarchy	**36.25**	36.44	**10.81**	12.05	0.19
Hier Two-Person VQ-VAE (ours)	36.32	**36.26**	10.86	**11.84**	**0.17**

Table 2. **Quantitative Evaluation of Dance Generation.** Comparison of motion generation quality between baselines, ablated versions, and our method on the DD100 dataset. **Bold** indicates best.

Method	FID ↓	Div →	PFID ↓	PFC ↓	CF (%) →	BAS ↑
GT	–	15.67	–	0.36	82.6	0.213
Duolando [Siyao et al. 2024]	12.21	14.72	13.18	16.22	74.7	0.202
GCD [Le et al. 2023a]	9.71	15.03	12.03	8.11	78.1	0.203
InterGen [Liang et al. 2024]	13.77	15.01	14.11	12.4	60.1	0.172
MoFusion [Dabral et al. 2023]	21.2	15.60	23.09	7.5	21.1	0.202
A1: w/o relational positioning	5.03	14.34	14.97	4.83	79.5	0.203
A2: w/o hier tokenization	4.77	14.45	15.44	5.88	80.2	0.197
A3: w/o unified representation	5.65	14.02	18.99	5.66	75.66	0.204
A4: w/o traj refinement	2.62	14.11	2.81	5.31	78.2	0.211
A5: 438-D music representation	2.45	14.99	2.87	3.45	72.1	0.193
A8: 3 level of hierarchy	1.55	15.62	2.95	5.45	70.1	0.210
DuetGen (ours)	**1.31**	**15.71**	**2.54**	**1.47**	**83.2**	**0.215**

5.2 User Study

We conducted a user study to perceptually evaluate our method against the three baselines. We presented each participant with 10 sets of anonymized animations. In each set, we showed two-person dance animations generated by the different methods, all from the same music, in randomized orders. We showed all the animations side-by-side, and asked participants to rate each animation on *motion quality*, *music-motion alignment*, and *partner coordination*, on 5-point Likert scales (more details in the appendix). 30 participants

from different age groups and genders, with different levels of expertise in dancing, took our study. Figure 4 summarizes the study results where we see DuetGen earning the highest ranking across all three evaluation aspects, consistently achieving scores above '4' (at least 22% higher than any other methods). At the other end, InterGen shows notably poor performance across all metrics (scoring below '2'), mainly due to its unnatural motion dynamics.

5.3 Qualitative Results

Figure 5 presents a qualitative comparison of dances generated by our method, the baselines, and the ablations. InterGen [Liang et al. 2024] exhibits significant spatial drift, with dancers gradually moving far apart. Duolando [Siyao et al. 2024] and GCD [Le et al. 2023a] show improved partner awareness, but still suffer from poor coordination and frequent interpenetration between dancers. Our ablations show that *w/o relational positioning* (A1) and *w/o unified representation* (A3), the dancers are mostly unaware of each other's motions, while *w/o hierarchical tokenization* (A2) fails to model the fine-grained end-effector movements in close interactions. In contrast, DuetGen produces well-coordinated dances, with natural synchronization and spatial coordination between the dancers throughout the sequence. Notably, when compared to ground-truth motions from the dataset, DuetGen demonstrates the ability to generate novel dance sequences rather than merely reproducing the learning data. We also compare the reconstruction quality of our hierarchical VQ-VAE with the ablations in Figure 6.

6 Conclusion

In this work, we present DuetGen, the first approach for music-driven, two-person dance motion synthesis. Our approach comprises two principal components: (1) a hierarchical VQ-VAE to encode two-person dance motions into discrete tokens at two levels of granularity, capturing both global motion semantics and fine-grained articulations, and (2) the two-stage transformers to generate these two-level tokens from music using generative masked modeling. Our method produces two-person dance motions with plausible interactions and precise music-motion alignment, outperforming baselines in both quantitative evaluations and user studies.

Limitations and Future Work. While our method demonstrates significant advancements, it has certain limitations. Our model does not explicitly address body shape variations, instead relying on ground-truth body shapes provided in the dataset. We currently do not model fine-grained finger interactions, primarily due to excessive noise in the captured finger motions in the DD100 dataset (Figure 7). Our model validation is also limited to the 1.9 hours of dance motion available in DD100, currently the only available duet dataset with diverse genres of dance styles and music. While collecting large-scale, high-quality duet dance datasets may be expensive, future works could explore learning from a combination of motion capture data and online dance videos to address the data scarcity.

Acknowledgments

This research was supported by Snap Inc., the EU Horizon 2020 grant Carousel+ (101017779), and the BMBF grant MOMENTUM (01IW22001).

References

Simon Alexanderson, Rajmund Nagy, Jonas Beskow, and Gustav Eje Henter. 2023. Listen, Denoise, Action! Audio-driven motion synthesis with diffusion models. *ACM Transactions on Graphics (TOG)* 42, 4 (2023), 1–20.

Okan Arikan and David A Forsyth. 2002. Interactive motion generation from examples. *ACM Transactions on Graphics (TOG)* 21, 3 (2002), 483–490.

Andreas Aristidou, Daniel Cohen-Or, Jessica K Hodgins, Yiorgos Chrysanthou, and Ariel Shamir. 2018. Deep motifs and motion signatures. *ACM Transactions on Graphics (TOG)* 37, 6 (2018), 1–13.

Nikos Athanasiou, Alpár Ceske, Markos Diomataris, Michael J. Black, and Gül Varol. 2024. MotionFix: Text-Driven 3D Human Motion Editing. In *SIGGRAPH Asia 2024 Conference Papers*.

Ho Yin Au, Jie Chen, Junkun Jiang, and Yike Guo. 2022. Choreograph: Music-conditioned automatic dance choreography over a style and tempo consistent dynamic graph. In *Proceedings of the 30th ACM International Conference on Multimedia*. 3917–3925.

Aneesh Bhattacharya, Manas Paranjape, Uttaran Bhattacharya, and Aniket Bera. 2024. DanceAnyWay: Synthesizing Beat-Guided 3D Dances with Randomized Temporal Contrastive Learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 783–791.

Zalán Borsos, Raphaël Marinier, Damien Vincent, Eugene Kharitonov, Olivier Pietquin, Matt Sharifi, Dominik Roblek, Olivier Teboul, David Grangier, Marco Tagliasacchi, et al. 2023. AudioLM: a language modeling approach to audio generation. *IEEE/ACM transactions on audio, speech, and language processing* 31 (2023), 2523–2533.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.

Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T Freeman. 2022. MaskGIT: Masked generative image transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 11315–11325.

Kang Chen, Zhipeng Tan, Jin Lei, Song-Hai Zhang, Yuan-Chen Guo, Weidong Zhang, and Shi-Min Hu. 2021. ChoreoMaster: choreography-oriented music-driven dance synthesis. *ACM Transactions on Graphics (TOG)* 40, 4 (2021), 1–13.

Baptiste Chopin, Hao Tang, Naima Otberdout, Mohamed Daoudi, and Nicu Sebe. 2023. Interaction Transformer for Human Reaction Generation. arXiv:2207.01685 [cs.CV] https://arxiv.org/abs/2207.01685

Rishabh Dabral, Muhammad Hamza Mughal, Vladislav Golyanik, and Christian Theobalt. 2023. MoFusion: A framework for denoising-diffusion-based motion synthesis. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

Ke Fan, Junshu Tang, Weijian Cao, Ran Yi, Moran Li, Jingyu Gong, Jiangning Zhang, Yabiao Wang, Chengjie Wang, and Lizhuang Ma. 2024. FreeMotion: A Unified Framework for Number-free Text-to-Motion Synthesis. arXiv:2405.15763 [cs.CV] https://arxiv.org/abs/2405.15763

Anindita Ghosh, Noshaba Cheema, Cennet Oguz, Christian Theobalt, and Philipp Slusallek. 2021. Text-based motion synthesis with a hierarchical two-stream rnn. In *ACM SIGGRAPH 2021 Posters*. 1–2.

Anindita Ghosh, Rishabh Dabral, Vladislav Golyanik, Christian Theobalt, and Philipp Slusallek. 2024. ReMoS: 3d motion-conditioned reaction synthesis for two-person interactions. In *European Conference on Computer Vision*. Springer, 418–437.

Kehong Gong, Dongze Lian, Heng Chang, Chuan Guo, Zihang Jiang, Xinxin Zuo, Michael Bi Mi, and Xinchao Wang. 2023. TM2D: Bimodality driven 3D dance generation via music-text integration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 9942–9952.

Chuan Guo, Yuxuan Mu, Muhammad Gohar Javed, Sen Wang, and Li Cheng. 2024a. Momask: Generative masked modeling of 3d human motions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1900–1910.

Chuan Guo, Yuxuan Mu, Xinxin Zuo, Peng Dai, Youliang Yan, Juwei Lu, and Li Cheng. 2024b. Generative Human Motion Stylization in Latent Space. In *The Twelfth International Conference on Learning Representations*.

Chuan Guo, Xinxin Zuo, Sen Wang, and Li Cheng. 2022b. TM2T: Stochastic and tokenized modeling for the reciprocal generation of 3d human motions and texts. In *European Conference on Computer Vision*.

Wen Guo, Xiaoyu Bie, Xavier Alameda-Pineda, and Francesc Moreno-Noguer. 2022a. Multi-person extreme motion prediction. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems* 30 (2017).

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems* 33 (2020).

Md Afzal Hossan, Sheeraz Memon, and Mark A Gregory. 2010. A novel approach for MFCC feature extraction. In *2010 4th international conference on signal processing and communication systems*. IEEE, 1–5.

Ruozi Huang, Huang Hu, Wei Wu, Kei Sawada, Mi Zhang, and Daxin Jiang. 2020. Dance revolution: Long-term dance generation with music via curriculum learning. *arXiv preprint arXiv:2006.06119* (2020).

Muhammad Gohar Javed, Chuan Guo, Li Cheng, and Xingyu Li. 2024. InterMask: 3D Human Interaction Generation via Collaborative Masked Modelling. *arXiv preprint arXiv:2410.10010* (2024).

Biao Jiang, Xin Chen, Wen Liu, Jingyi Yu, Gang Yu, and Tao Chen. 2023. MotionGPT: Human motion as a foreign language. *Advances in Neural Information Processing Systems* 36 (2023), 20067–20079.

Nhat Le, Tuong Do, Khoa Do, Hien Nguyen, Erman Tjiputra, Quang D Tran, and Anh Nguyen. 2023a. Controllable group choreography using contrastive diffusion. *ACM Transactions on Graphics (TOG)* 42, 6 (2023), 1–14.

Nhat Le, Thang Pham, Tuong Do, Erman Tjiputra, Quang D Tran, and Anh Nguyen. 2023b. Music-driven group choreography. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8673–8682.

Hsin-Ying Lee, Xiaodong Yang, Ming-Yu Liu, Ting-Chun Wang, Yu-Ding Lu, Ming-Hsuan Yang, and Jan Kautz. 2019. Dancing to Music. In *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Eds.), Vol. 32. Curran Associates, Inc.

Buyu Li, Yongchi Zhao, Shi Zhelun, and Lu Sheng. 2022. Danceformer: Music conditioned 3d dance generation with parametric motion transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 1272–1279.

Jiaman Li, Yihang Yin, Hang Chu, Yi Zhou, Tingwu Wang, Sanja Fidler, and Hao Li. 2020. Learning to generate diverse dance motions with transformer. *arXiv preprint arXiv:2008.08171* (2020).

Ruilong Li, Shan Yang, David A Ross, and Angjoo Kanazawa. 2021. Ai choreographer: Music conditioned 3d dance generation with aist++. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 13401–13412.

Ronghui Li, YuXiang Zhang, Yachao Zhang, Hongwen Zhang, Jie Guo, Yan Zhang, Yebin Liu, and Xiu Li. 2024a. Lodge: A coarse to fine diffusion network for long dance generation guided by the characteristic dance primitives. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1524–1534.

Ronghui Li, Youliang Zhang, Yachao Zhang, Yuxiang Zhang, Mingyang Su, Jie Guo, Ziwei Liu, Yebin Liu, and Xiu Li. 2024b. InterDance: Reactive 3D Dance Generation with Realistic Duet Interactions. *arXiv preprint arXiv:2412.16982* (2024).

Ronghui Li, Junfan Zhao, Yachao Zhang, Mingyang Su, Zeping Ren, Han Zhang, Yansong Tang, and Xiu Li. 2023. FineDance: A fine-grained choreography dataset for 3d full body dance generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 10234–10243.

Zhengyuan Li, Kai Cheng, Anindita Ghosh, Uttaran Bhattacharya, Liangyan Gui, and Aniket Bera. 2025. SimMotionEdit: Text-Based Human Motion Editing with Motion Similarity Prediction. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2025).

Han Liang, Wenqian Zhang, Wenxuan Li, Jingyi Yu, and Lan Xu. 2024. Intergen: Diffusion-based multi-human motion generation under complex interactions. *International Journal of Computer Vision* (2024), 1–21.

Yunze Liu, Changxi Chen, and Li Yi. 2023. Interactive humanoid: Online full-body motion reaction synthesis with social affordance canonicalization and forecasting. *arXiv preprint arXiv:2312.08983* (2023).

Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. 2023. SMPL: A skinned multi-person linear model. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*. 851–866.

Vongani H Maluleke, Lea Müller, Jathushan Rajasegaran, Georgios Pavlakos, Shiry Ginosar, Angjoo Kanazawa, and Jitendra Malik. 2024. Synergy and Synchrony in Couple Dances. *arXiv preprint arXiv:2409.04440* (2024).

Evonne Ng, Hanbyul Joo, Liwen Hu, Hao Li, Trevor Darrell, Angjoo Kanazawa, and Shiry Ginosar. 2022. Learning to listen: Modeling non-deterministic dyadic facial motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 20395–20405.

Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. 2019. Expressive Body Capture: 3D Hands, Face, and Body from a Single Image. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. doi:10.1109/CVPR.2019.01123

Mathis Petrovich, Michael J Black, and Gül Varol. 2022. TEMOS: Generating diverse human motions from textual descriptions. In *European Conference on Computer Vision*.

Pablo Ruiz Ponce, German Barquero, Cristina Palmero, Sergio Escalera, and Jose Garcia-Rodriguez. 2024. in2IN: Leveraging individual Information to Generate Human INteractions. arXiv:2404.09988 [cs.CV] https://arxiv.org/abs/2404.09988

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PMLR, 8748–8763.

Katerina El Raheb, Marina Stergiou, Akrivi Katifori, and Yannis Ioannidis. 2019. Dance Interactive Learning Systems: A Study on Interaction Workflow and Teaching Approaches. *ACM Comput. Surv.* 52, 3, Article 50 (June 2019), 37 pages.

Yonatan Shafir, Guy Tevet, Roy Kapon, and Amit H. Bermano. 2023. Human Motion Diffusion as a Generative Prior. arXiv:2303.01418 [cs.CV] https://arxiv.org/abs/2303.01418

Ayush Shah, Manasi Kattel, Araju Nepal, and D Shrestha. 2019. Chroma feature extraction. *Chroma Feature Extraction using Fourier Transform* (2019).

Mengyi Shan, Lu Dong, Yutao Han, Yuan Yao, Tao Liu, Ifeoma Nwogu, Guo-Jun Qi, and Mitch Hill. 2024. Towards Open Domain Text-Driven Synthesis of Multi-Person Motions. arXiv:2405.18483 [cs.CV] https://arxiv.org/abs/2405.18483

Takaaki Shiratori, Atsushi Nakazawa, and Katsushi Ikeuchi. 2006. Dancing-to-music character animation. In *Computer Graphics Forum*, Vol. 25. Wiley Online Library, 449–458.

Li Siyao, Tianpei Gu, Zhitao Yang, Zhengyu Lin, Ziwei Liu, Henghui Ding, Lei Yang, and Chen Change Loy. 2024. Duolando: Follower GPT with Off-Policy Reinforcement Learning for Dance Accompaniment. In *International Conference on Learning Representations (ICLR)*.

Li Siyao, Weijiang Yu, Tianpei Gu, Chunze Lin, Quan Wang, Chen Qian, Chen Change Loy, and Ziwei Liu. 2022. Bailando: 3d dance generation by actor-critic gpt with choreographic memory. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 11050–11059.

Sebastian Starke, Yiwei Zhao, Taku Komura, and Kazi Zaman. 2020. Local motion phases for learning multi-contact character movements. *ACM Transactions on Graphics (TOG)* (2020).

Sebastian Starke, Yiwei Zhao, Fabio Zinno, and Taku Komura. 2021. Neural animation layering for synthesizing martial arts movements. *ACM Transactions on Graphics (TOG)* 40, 4 (2021), 1–16.

Kewei Sui, Anindita Ghosh, Inwoo Hwang, Jian Wang, and Chuan Guo. 2025. A Survey on Human Interaction Motion Generation. *arXiv preprint arXiv:2503.12763* (2025).

Guofei Sun, Yongkang Wong, Zhiyong Cheng, Mohan S Kankanhalli, Weidong Geng, and Xiangdong Li. 2020. Deepdance: music-to-dance motion choreography with adversarial learning. *IEEE Transactions on Multimedia* 23 (2020), 497–509.

Jiangxin Sun, Chunyu Wang, Huang Hu, Hanjiang Lai, Zhi Jin, and Jian-Fang Hu. 2022. You never stop dancing: Non-freezing dance generation via bank-constrained manifold projection. *Advances in Neural Information Processing Systems* 35 (2022), 9995–10007.

Jonathan Tseng, Rodrigo Castellon, and Karen Liu. 2023. EDGE: Editable dance generation from music. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Aaron Van Den Oord, Oriol Vinyals, et al. 2017. Neural discrete representation learning. *Advances in neural information processing systems* 30 (2017).

Zixuan Wang, Jia Jia, Haozhe Wu, Junliang Xing, Jinghe Cai, Fanbo Meng, Guowen Chen, and Yanfeng Wang. 2022. Groupdancer: Music to multi-people dance synthesis with style collaboration. In *Proceedings of the 30th ACM International Conference on Multimedia*. 1138–1146.

Zhenzhi Wang, Jingbo Wang, Yixuan Li, Dahua Lin, and Bo Dai. 2024. InterControl: Zero-shot Human Interaction Generation by Controlling Every Joint. arXiv:2311.15864 [cs.CV] https://arxiv.org/abs/2311.15864

Liang Xu, Xintao Lv, Yichao Yan, Xin Jin, Shuwen Wu, Congsheng Xu, Yifan Liu, Yizhou Zhou, Fengyun Rao, Xingdong Sheng, et al. 2024a. Inter-x: Towards versatile human-human interaction analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 22260–22271.

Liang Xu, Yizhou Zhou, Yichao Yan, Xin Jin, Wenhan Zhu, Fengyun Rao, Xiaokang Yang, and Wenjun Zeng. 2024b. ReGenNet: Towards Human Action-Reaction Synthesis. arXiv:2403.11882 [cs.CV] https://arxiv.org/abs/2403.11882

Siyue Yao, Mingjie Sun, Bingliang Li, Fengyu Yang, Junle Wang, and Ruimao Zhang. 2023. Dance with you: The diversity controllable dancer generation via diffusion models. In *Proceedings of the 31st ACM International Conference on Multimedia*. 8504–8514.

Canyu Zhang, Youbao Tang, Ning Zhang, Ruei-Sung Lin, Mei Han, Jing Xiao, and Song Wang. 2024b. Bidirectional Autoregessive Diffusion Model for Dance Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 687–696.

Jianrong Zhang, Yangsong Zhang, Xiaodong Cun, Yong Zhang, Hongwei Zhao, Hongtao Lu, Xi Shen, and Ying Shan. 2023. T2M-GPT: Generating Human Motion From Textual Descriptions With Discrete Representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Yaqi Zhang, Di Huang, Bin Liu, Shixiang Tang, Yan Lu, Lu Chen, Lei Bai, Qi Chu, Nenghai Yu, and Wanli Ouyang. 2024a. MotionGPT: Finetuned llms are general-purpose motion generators. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 7368–7376.

Zixiang Zhou and Baoyuan Wang. 2023. UDE: A unified driving engine for human motion generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5632–5641.

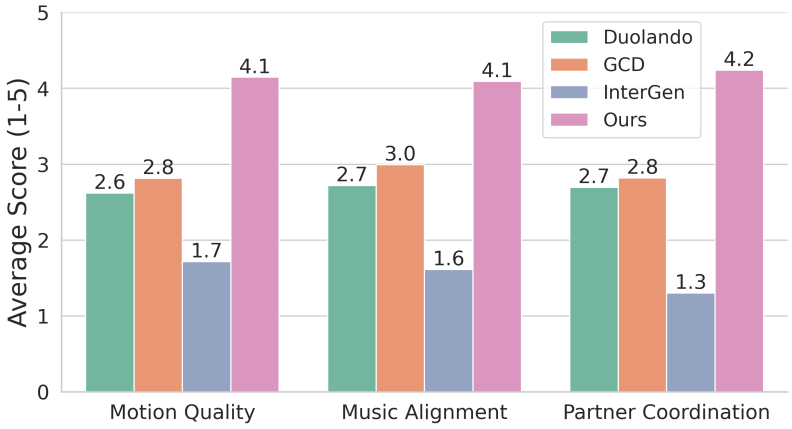

Fig. 4. **User Study Results.** Each column indicates the average user rating on a 1-5 scale. DuetGen consistently outperforms all baselines.

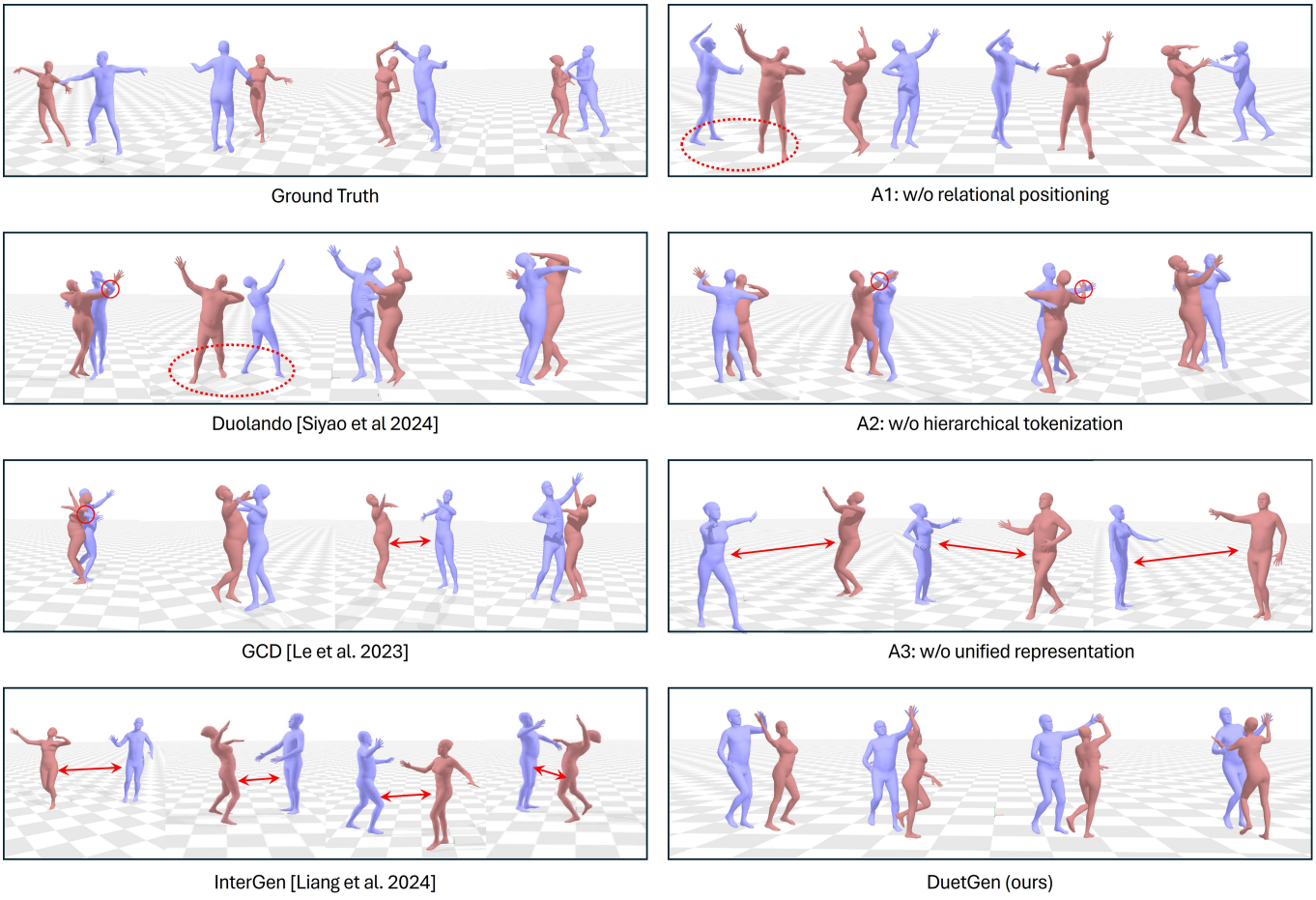

Fig. 5. **Qualitative Comparisons.** Dance motions generated by DuetGen, the baselines, and relevant ablations, from the same music input. Notice that the baseline methods and the ablations exhibit uncoordinated movements (*red dots*), interpenetration (*red circles*), and drift in root joint positions (*red arrows*). In contrast, DuetGen maintains natural interactions and well-synchronized two-person dance movements.

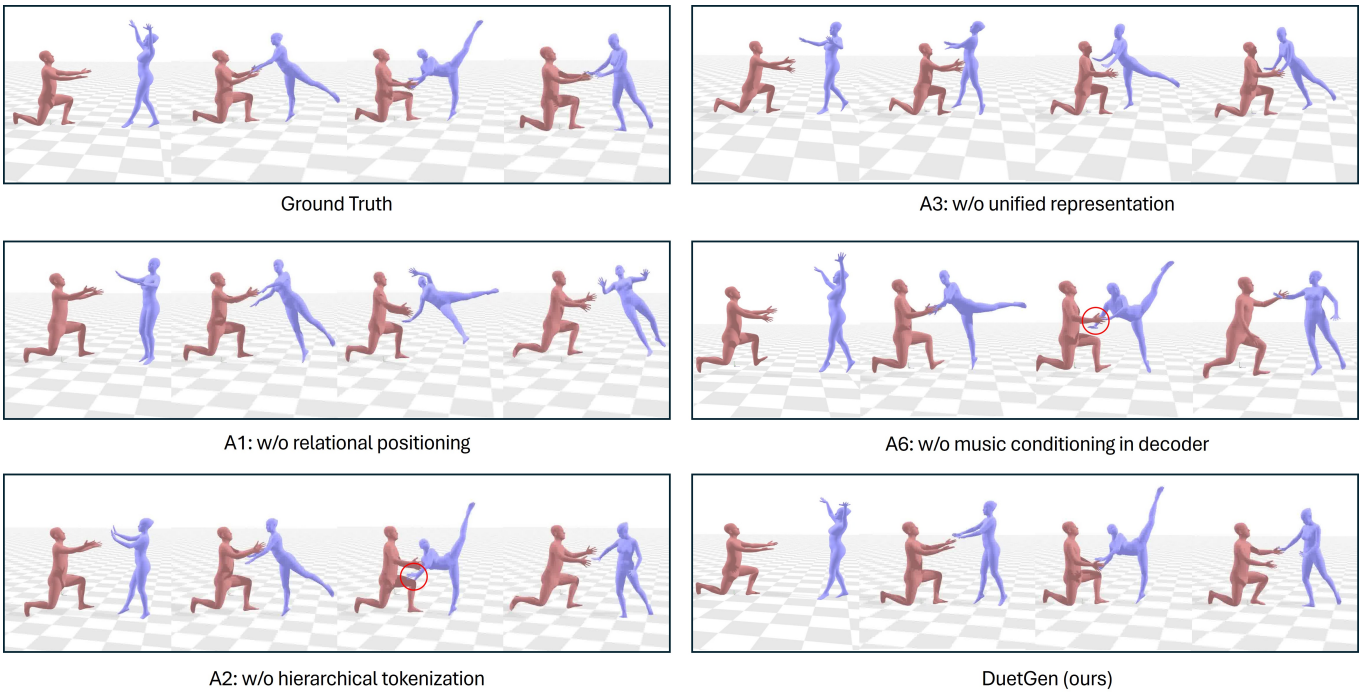

Fig. 6. **Qualitative Comparison on VQ-VAE reconstruction.** Reconstruction quality of the hierarchical two-person VQ-VAE module of DuetGen compared to its ablations. Notice that the ablations exhibit uncoordinated movements and interpenetration (*red circles*), while DuetGen achieves interactions and synchronization between the two persons.

Fig. 7. **Noisy Finger Motions in the DD100 dataset.** Common artifacts in the DD100 dataset include twisted or inter-penetrated finger motions (*red circles*).