EgoMusic-driven Human Dance Motion Estimation with Skeleton Mamba

Quang Nguyen[1], Nhat Le[2], Baoru Huang[7,*], Minh Nhat Vu[3], Chengcheng Tang[4], Van Nguyen[1], Ngan Le[5], Thieu Vo[6], Anh Nguyen[7]

[1]FPT Software AI Center [2]The University of Western Australia [3]TU Wien [4]Meta

[5]University of Arkansas [6]National University of Singapore [7]University of Liverpool [*]Corresponding author

https://zquang2202.github.io/SkeletonMamba/

Figure 1. We present a new dataset and method for estimating human dance motion from the egocentric video and music.

Abstract

Estimating human dance motion is a challenging task with various industrial applications. Recently, many efforts have focused on predicting human dance motion using either egocentric video or music as input. However, the task of jointly estimating human motion from both egocentric video and music remains largely unexplored. In this paper, we aim to develop a new method that predicts human dance motion from both egocentric video and music. In practice, the egocentric view often obscures much of the body, making accurate full-pose estimation challenging. Additionally, incorporating music requires the generated head and body movements to align well with both visual and musical inputs. We first introduce EgoAIST++, a new large-scale dataset that combines both egocentric views and music with more than 36 hours of dancing motion. Drawing on the success of diffusion models and Mamba on modeling sequences, we develop an EgoMusic Motion Network with a core Skeleton Mamba that explicitly captures the skeleton structure of the human body. We illustrate that our approach is theoretically supportive. Intensive experiments show that our method clearly outperforms state-of-the-art approaches and generalizes effectively to real-world data.

1. Introduction

Dance is a fundamental form of human expression, creativity, and is deeply embedded in cultural and social contexts [21, 36]. Estimating full-body dance motion is a crucial task with many industrial applications, such as dance education [13, 20, 38], virtual metaverses [33, 35], or film animation [6, 84]. While several works have focused on human dance pose estimation, they mostly tackle the problem using the input from third-person video [4, 25, 64, 65, 98] or motion tracking device [56, 57]. In practice, third-person video methods suffer from occlusions, viewpoint variations, and depth ambiguity, while motion-tracking devices require costly hardware, making them less practical for real-world AR/VR or metaverse applications. These constraints highlight the need for an alternative approach, such as first-person (egocentric) view dance motion estimation.

Recently, egocentric input has been utilized to estimate human motions in everyday activities such as walking and running [30, 59, 80, 81]. A key challenge in egocentric pose estimation is that much of the body often falls outside the camera's view, creating ambiguities in full-body capture. This issue is even more pronounced in dance motion, where movements are more complex and dynamic. To overcome this, many works incorporate motion tracking

sensors, which require costly hardware and limit accessibility [56, 57]. A promising approach for accurate dance motion estimation is to leverage music as an additional modality. Research has shown that music, through its rhythm, tempo, and dynamics, can provide meaningful cues for generating realistic movements [5, 36]. Therefore, we propose integrating *music and egocentric video* for dance motion estimation. We hypothesize that combining these two complementary modalities enables more accurate human dance motion estimation. However, the challenge lies in two fundamental tasks: *i)* creating a large-scale dataset for dance motion estimation from egocentric and music, and *ii)* designing a model capable of understanding the human skeleton structure and effectively coordinating multimodal inputs to synchronize head and body motion with the music

To learn human motion, transformer is a widely used technique [8, 45, 74, 77, 94]. However, transformer has quadratic complexity and struggles to capture structure dependencies. On the other hand, State Space Models (or Mamba) [10,18] have shown great potential on several tasks over transformer-based models, such as graph analysis [79], video analysis [44], and image generation [24]. However, directly adapting Mamba to human motion data presents a challenge due to the dynamics of both spatial and temporal structures of the human body. Previous Mamba-based models [63, 85, 95, 96] on human motion usually simplify each frame of a human pose as a single latent vector or disregard the spatial order of joints within the human skeleton, which limits their ability to capture fine-grained spatial dynamics. This drawback makes it challenging to generate coherent motion as the head and lower body may fail to align naturally, leading to poor coordination between egocentric-driven head and music-influenced body movement.

In this work, we first introduce a new dataset for human dance pose estimation from egocentric and music inputs. We then propose Skeleton Mamba, a new Mamba model designed to capture spatial structures while preserving temporal coherence. Our designed method enables synchronized head and body movements responsive to egocentric and music inputs. Our approach explicitly models the spatial structure of joints and their hierarchical dependencies, allowing for more coherent motion generation that preserves the natural relationships between joints. We show that our method is theoretically supportive, and provide intensive experiments to validate our method against recent state-of-the-art approaches. Our contributions are the following:

- We propose a new dataset for human dance motion estimation from the egocentric and music input.
- We propose the EgoMusic Motion Network with Skeleton Mamba as the core to learn human body motion.
- We provide theoretical analysis and intensive experiments to demonstrate the effectiveness of our method.

2. Related Work

Human Motion from Egocentric Video. Human motion estimation from egocentric video has garnered significant attention in recent years. Most existing methods assume partial visibility of body parts in the image, often using fisheye cameras [30,59,75,80,81,87]. Other research addresses the challenge of body parts not being visible in egocentric footage [29,43,55,60]. Jiang *et al.* [29] introduce an innovative global optimization technique that utilizes both trained dynamic and scene classifiers along with pose coupling over an extended period. Ng *et al.* [60] model person-to-person interactions, inferring the 3D ego-pose based on the other person's pose. Luo *et al.* [55] jointly models kinematics and dynamics to estimate 3D human poses and human-object interactions. EgoFormer [47] extracts motion features from egocentric images and employs a Transformer Decoder to autoregressively generate human poses. In [43], the authors introduce EgoEgo, a hybrid learning method for head pose estimation, which then was used as a conditioning factor in a diffusion model to estimate full-body motions.

Human Motion from Music. Generating human dance motion from Music is widely formed as a synthesis task. Early studies used statistical retrieval techniques to generate choreography by seamlessly transitioning between existing motion clips [16, 41]. However, these methods rely on selecting pre-existing motions, often resulting in unnatural dance motions. With advancements in deep learning techniques and the availability of large-scale datasets, many networks have been introduced to generate higher-fidelity dance motions [2, 15, 27, 37, 39, 42, 45, 69]. The FACT model [45] introduces an autoregressive cross-modal transformer to generate long continuous dance sequences. Bailando [69] employs VQ-VAEs for the upper and lower body segments to translate music and initial poses into dance sequences. Recent efforts have explored the use of diffusion models for dance generation [45, 92]. MoFusion [9] presents a multi-condition diffusion framework capable of generating long, realistic, and temporally coherent human motion sequences. EDGE [77] introduces an editable dance generation model that leverages a transformer-based diffusion architecture, offering flexible editing capabilities for dance applications. In [28], the authors propose a framework that allows control of generated dance motion based on key-frame body pose and music beat conditions. However, all these works focus on generating dance motion using only the music as input, and the egocentric views are not taken into account. In this work, we present a new diffusion framework that aligns body movements with both music and egocentric video.

State Space Model. State Space Model (SSM) [18] has garnered significant attention recently due to its potential for efficiently modeling long sequences with linear complexity. Its applications have been explored across vari-

ous fields, including image processing [24, 40, 46, 86, 99], graph processing [3, 79], point cloud analysis [48, 50, 93], and human motion generation [96]. Vim [99] presents a bidirectional SSM block. Efforts like Mamba-ND [46] extend the capabilities of SSM to higher-dimensional data by exploring different scan directions within a single SSM block. In addition, several works, including ZigMa [24] and DiffuSSM [89], utilize Mamba-based SSM blocks for efficient image generation. MotionMamba [96] proposes a symmetric multi-branch Mamba that processes temporal and spatial and shows exceptional performance on text-to-motion generation tasks. More recently, State Space Duality (SSD) [10] is introduced as a dual-form framework that unifies state space models with structured masked attention.

3. The EgoAIST++ Dataset

While several datasets have been proposed for single input (either egocentric or music) human motion estimation (Table 1), large-scale datasets that combine egocentric and music for dance pose are still limited. Ego-Exo4D dataset [17] has a subset with the egocentric view, music, and dance motion, but this subset is only approximately 2 hours. To address this gap, we introduce EgoAIST++, a new large-scale dataset that integrates egocentric views and music specifically for human dance pose estimation.

Datasets	Music	Egocentric	Setup	#Images	Camera Direction
Mo^{2}Cap^2 [87]	✗	✓	Mocap	530k	Downward-facing
xr-EgoPose [75]	✗	✓	Simulation	380k	Downward-facing
UnrealEgo [1]	✗	✓	Simulation	450k	Downward-facing
EgoGTA [82]	✗	✓	Simulation	320K	Downward-facing
EgoBody3M [97]	✗	✓	Mocap	3.4M	Downward-facing
ECHP [52]	✗	✓	Mocap	75k	Downward-facing
ARES [43]	✗	✓	Simulation	1.6M	Forward-facing
Ego-Exo4D [17]	✓	✓	Mocap	/	Forward-facing
DanceNet [101]	✓	✗	-	-	-
EA-MUD [71]	✓	✗	-	-	-
AIST++ [45]	✓	✗	-	-	-
EgoAIST++ (ours)	✓	✓	Mixed	3.9M	Forward-facing

Table 1. **Human motion datasets comparison.**

Setup. We first utilize the AIST++ dataset [45] as it includes real-world well-defined human motion paired with the music. The data from AIST++ dataset was captured using a motion capture system which ensures the correctness of the human motion. We then use the Replica 3D indoor scene dataset [70] to provide environment for obtaining the visual egocentric view. We randomly placed AIST++ [45] sequences with a specified location and rotation in the 3D mesh scene of the Replica dataset. For each sequence, we enforce the penetration constraint as in [83] to maintain the natural contact between the human and the objects in the scene. Other factors such as collision with surrounding objects are resolved manually by human annotators.

Data Labelling and Statistic. We use AI Habitat [72], to render high-quality and realistic egocentric images from a head-mounted camera of a virtual human and 3D mesh scene. We split the data from the AIST++ and Replica datasets to ensure the train and test sets have distinct music choreographies and scenes, with no overlap between them. The test set includes 40 unique music choreographies and 5 distinct scenes, while the training set consists of 980 music choreographies and 13 scenes for training. For each dance sequence, we divide it into 5-second subsequences and place them at a random location within the scene. Overall, our EgoAIST++ dataset has 36 hours of motion with nearly 3.9M frames, recorded at 30 frames per second.

4. EgoMusic-driven Dance Motion Estimation

4.1. Problem Formulation

Given an egocentric video represented as a sequence of frames, $\mathbf{v} = \{\mathbf{v}^1, \mathbf{v}^2, \dots, \mathbf{v}^T\}$, and a piece of music, $\mathbf{a} = \{\mathbf{a}^1, \mathbf{a}^2, \dots, \mathbf{a}^T\}$, both of duration T, our objective is to generate a human dance sequence, $\mathbf{x} = \{\mathbf{x}^1, \mathbf{x}^2, \dots, \mathbf{x}^T\}$, that aligns with the egocentric view and the audio. As in [43, 77], the human pose $\mathbf{x}$ is presented using the SMPL model [54]. We formulate our problem using a condition diffusion model [12, 23] where we represent the target motion as $\mathbf{x}_0$, and combine egocentric images $\mathbf{v}$ and music $\mathbf{a}$ as the condition $\mathbf{z}$ in the diffusion model. The objective of the diffusion process [23] is to gradually add noise into a clean dance motion $\mathbf{x}_0$ over a series of m steps:

$$q(\mathbf{x}_m|\mathbf{x}_0) = \mathcal{N}(\sqrt{\alpha_m}\mathbf{x}_0, (1-\alpha_m)\mathbf{I}) \,, \tag{1}$$

where $\alpha_m = \prod_{s=1}^{m}(1-\beta_s)$, and β_m controls the noise schedule. The backward process is to learn the condition distribution $p_\theta(\mathbf{x}_0|\mathbf{z})$ with the condition $\mathbf{z}$ as in [12, 23].

Our objective is to design an effective backward process. To generate human motion that is well-aligned with egocentric cues and music, we introduce a new EgoMusic Motion Network (EMM) centered around a core Skeleton Mamba scanning strategy to learn the structure of the human body. We provide empirical and theoretical evidence demonstrating the contribution of our Skeleton Mamba in learning human motion during the denoising process.

4.2. EgoMusic Motion Network

The overall pipeline of our proposed method is illustrated in Fig. 2. First, we extract the features from the input egocentric images $\mathbf{v}$ using a deep network [22]. For music $\mathbf{a}$, we adopt the same feature extraction process as EDGE [77], leveraging the pre-trained JukeBox [11] model to capture high-level music features, which are then processed by a transformer encoder [78] to produce the final music embedding. The visual and music embeddings are subsequently aligned and integrated by the Fusion Module

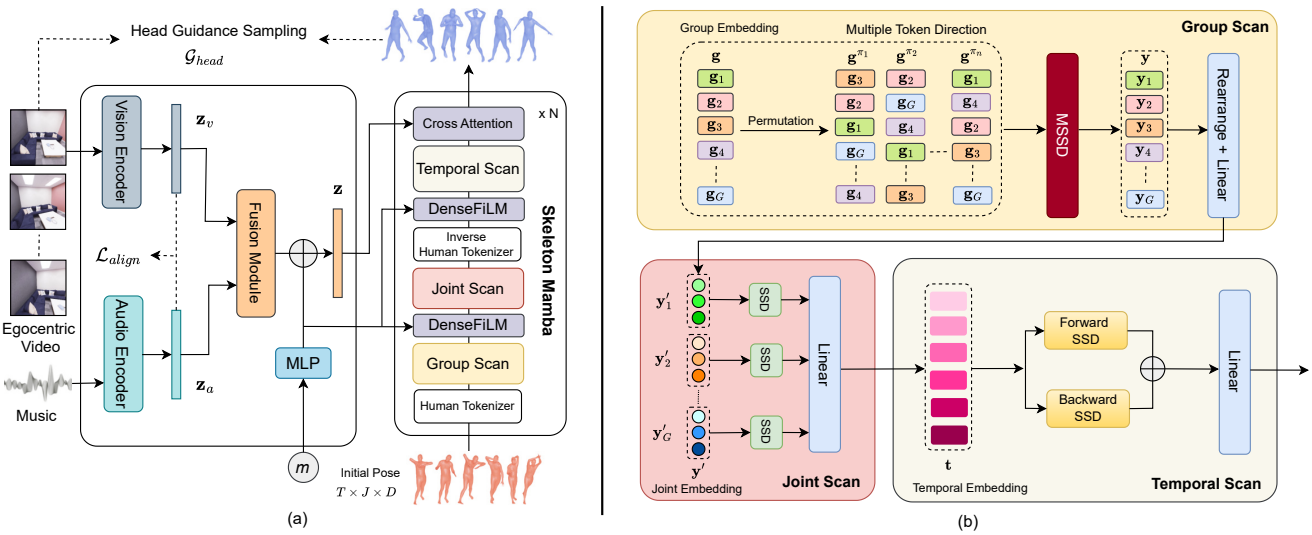


Figure 2. **Methodology overview.** (a) We propose a new diffusion model framework that generates human motion from egocentric video and music. (b) Detail architecture of three main components: Group Scan, Joint Scan, and Temporal Scan. Our model can effectively capture both the spatial and temporal dynamics in human motion data.

to form a joined embedding, denoted as $\mathbf{z}$. This embedding is then fed into a conditional diffusion denoising process, which outputs the final denoised dance motion sequence. The denoising process is guided by the proposed Skeleton Mamba, which is designed to maintain the skeletal structure of the human pose and enhance the smoothness of the generated motion. We use feature-wise linear DenseFiLM [61] for timestep encoding and a Cross Attention layer to integrate the condition $\mathbf{z}$ into the denoising process.

4.3. Skeleton Mamba

Motivation. Estimating human pose from egocentric video and music requires spatial, temporal, and visual understanding. Previous methods [43, 77, 96] for human motion generation often overlook the structured patterns inherent in the human body. Skeleton-based methods, widely used in action recognition tasks [62, 90], effectively capture essential motion dynamics while handling challenges like partial visibility and occlusions common in egocentric perspectives. These methods show potential for our task, where a semantic understanding of body part dynamics is crucial. To address these needs, we propose Skeleton Mamba, a model that learns detailed body structures to enhance pose accuracy, ensuring head alignment with the egocentric view and synchronizing body movements with musical cues. Our Skeleton Mamba includes the Group Scan and Joint Scan strategy to learn the group-level representation (e.g., left arm, right arm) and joint-level dependencies within the human body. Then a Temporal Scan is applied to capture sequential dependencies over time. Our Skeleton Mamba is thus capable of effectively modeling both spatial and temporal dynamics within human motion data.

Human Tokenizer. Given the human representation $\mathbf{x} \in \mathbb{R}^{T \times J \times D}$, where T is the number of frames, J is the number of joints, and D is the dimension, we first tokenize the human pose into G overlapping joint groups, each group containing P joints. This results in a group sequence represented as $\mathbf{g} \in \mathbb{R}^{T \times G \times E}$:

$$\mathbf{g} = [\mathbf{g}_1, \mathbf{g}_2, \ldots, \mathbf{g}_G] = \text{HumanTokenizer}(\mathbf{x}) \,, \quad (2)$$

where $\mathbf{g}_i \in \mathbb{R}^{T \times 1 \times E}$ represents the embedding of each token in the group sequence, and $E = P \times D$ is the embedding dimension. The joints of the human body are grouped based on skeletal parts, with some joints shared across multiple groups. This grouping strategy reflects the symmetrical structure of the human body.

Group Scan with Multi-directional SSD. Inspired by recent works [85, 96] that employed State Space Model (SSM) [18] for human motion data, we extend these ideas by adopting a multi-directional approach within the grouped structure, leveraging the State Space Duality (SSD) [10] to encode human motion at the group level. Our Group Scan rearranges the group tokens in multiple ways and learns them from multiple directions to enable a more comprehensive exploration of token relationships. Algorithm 1 shows our Multi-directional SSD (MSSD). We first utilize n permutation operators $[\pi_1, \pi_2, \ldots, \pi_n]$, where each π_i is the element of symmetric group $\text{Sym}(G)$. Each π_i reorders the group tokens of the group embedding $\mathbf{g} \in \mathbb{R}^{T \times G \times E}$. For each permutation π_i, the reordered embedding $\mathbf{g}_i$ is obtained as $\mathbf{g}^{\pi_i} = [\mathbf{g}_{\pi_i(1)}, \ldots, \mathbf{g}_{\pi_i(G)}] \in \mathbb{R}^{T \times G \times E}$. These n reordered embeddings are concatenated along the group dimension, forming a combined representation $\bar{\mathbf{g}}_c \in \mathbb{R}^{T \times nG \times E}$. Embedding $\bar{\mathbf{g}}_c$ is processed by

an SSD, and the resulting output is split back into n segments: $[\bar{\mathbf{g}}^{\pi_1}, \bar{\mathbf{g}}^{\pi_2}, \ldots, \bar{\mathbf{g}}^{\pi_n}]$ with $\bar{\mathbf{g}}^{\pi_i} \in \mathbb{R}^{T\times G\times E}$. Each segment is then reordered back to its original token order using the corresponding inverse permutation π_i^{-1}. Finally, the mean of the reordered embeddings is taken to produce the transformed group embedding $\mathbf{y} \in \mathbb{R}^{T\times G\times E}$, where $\mathbf{y}_i \in \mathbb{R}^{T\times 1\times E}$ represents the i-th token in the transformed group sequence.

Algorithm 1 Multi-directional SSD (MSSD)

Input: Group Embedding $\mathbf{g} : (T, G, E)$;
n permutation $(\pi_1, \pi_2, \ldots, \pi_n), \pi_i \in \mathrm{Sym}(G)$.
Output: Transformed group embedding $\mathbf{y} : (T, G, E)$.

for $i = 1, \ldots, n$ **do**
  / permutation sequence */*
  $\mathbf{g}^{\pi_i} : (T, G, E) = [\mathbf{g}_{\pi_i(1)}, \ldots, \mathbf{g}_{\pi_i(G)}]$
end for
$\mathbf{g}_c : (T, nG, E) \leftarrow \mathrm{Concat}([\mathbf{g}^{\pi_1}, \mathbf{g}^{\pi_2}, ..., \mathbf{g}^{\pi_n}])$
$\bar{\mathbf{g}}_c : (T, nG, E) \leftarrow \mathrm{SSD}(\mathbf{g}_c)$
$[\bar{\mathbf{g}}^{\pi_1}, \bar{\mathbf{g}}^{\pi_2}, \ldots, \bar{\mathbf{g}}^{\pi_n}] \leftarrow \mathrm{Split}(\bar{\mathbf{g}}_c)$
for $i = 1, \ldots, n$ **do**
  / reverse to the original order */*
  $\bar{\mathbf{g}}^{\pi_i^{-1}} : (T, G, E) = [\bar{\mathbf{g}}_{\pi_i^{-1}(1)}, \ldots, \bar{\mathbf{g}}_{\pi_i^{-1}(G)}]$
end for
$\mathbf{y} : (T, G, E) \leftarrow \mathrm{Mean}([\bar{\mathbf{g}}^{\pi_1^{-1}}, \bar{\mathbf{g}}^{\pi_2^{-1}}, \ldots, \bar{\mathbf{g}}^{\pi_n^{-1}}])$
Return $\mathbf{y}$

Joint Scan. To learn the human motion at the joint level, we transform each group embedding $\mathbf{y}_i$ to a sequence of individual joints represented as $\mathbf{y}'_i \in \mathbb{R}^{T\times P\times D}$ using a linear layer and rearrange operator. Each of these sequences is then processed by an SSD block to obtain transformed joint sequence embedding $\mathbf{y}''_i \in \mathbb{R}^{T\times P\times D}$:

$$\begin{aligned} \mathbf{y}'_i &= \mathrm{Rearrange}(\mathrm{Linear}(\mathbf{y}_i)) \,, \\ \mathbf{y}''_i &= \mathrm{SSD}(\mathbf{y}'_i) \,. \end{aligned} \tag{3}$$

In total, G separate SSD modules with shared weights are employed. The objective is to learn detailed intra-group dependencies, ensuring that the interactions within each joint sequence are effectively captured. Unlike group tokens, joint tokens have an inherent order, allowing a unidirectional SSD to be sufficient for learning. By integrating the Group Scan and Joint Scan, our approach effectively captures both high-level and detailed spatial dependencies in the human pose structure. All outputs of the SSD modules are then concatenated and fed to the Inverse Human Tokenizer to restore the original pose shape $\mathbf{t} \in \mathbb{R}^{T\times J\times D}$.

$$\mathbf{t} = \mathrm{InverseHumanTokenizer}((\mathrm{Concat}(\mathbf{y}''_1, \mathbf{y}''_2, \ldots, \mathbf{y}''_G))) \,. \tag{4}$$

Temporal Scan. We apply Temporal Scan to model the temporal dependencies across the temporal domain, enhancing the representation of dynamic pose changes over time. First, we swap the input dimension from $\mathbf{t} \in \mathbb{R}^{T\times J\times D}$ to $\mathbf{t}' \in \mathbb{R}^{J\times T\times D}$. Then, the embedding $\mathbf{t}'$ is processed by two SSD modules that scan the sequence in both backward and forward directions as in [44, 85], producing $\mathbf{t}_{backward}$ and $\mathbf{t}_{forward}$. Unlike Group Scan, which uses multiple directions, temporal sequences must preserve natural dependencies over time, allowing only backward and forward scanning.

$$\begin{aligned} \mathbf{t}_{backward} &= \mathrm{BackwardSSD}(\mathbf{t}') \,, \\ \mathbf{t}_{forward} &= \mathrm{ForwardSSD}(\mathbf{t}') \,. \end{aligned} \tag{5}$$

Theorem 1. *Let $S_J = Sym(J)$ denote the symmetric group of J elements, and let HT, HT^{-1} be short for the Human-Tokenizer and InverseHumanTokenizer respectively. Suppose that the $HT(\cdot)$ and $HT^{-1}(\cdot)$ operator are fixed, such that the set $H = \{\sigma \in S_J | HT(\mathbf{x}^\sigma) = HT(\mathbf{x})\}$ is a non-empty subgroup of S_J. Then for arbitrary continuous and H-equivariant function $g : \mathbb{R}^{J\times D} \to \mathbb{R}^{J\times D}$, compact set $K \subseteq \mathbb{R}^{J\times D}$, and $\epsilon > 0$, there exists a function $f : \mathbb{R}^{J\times D} \to \mathbb{R}^{J\times D}$ that constructed by our Skeleton Mamba such that:*

$$||f(\mathbf{x}) - g(\mathbf{x})||_\infty < \epsilon, \ \forall \mathbf{x} \in K.$$

Proof. See Supplementary Material. □

Remark 1.1. The group H in the above theorem represents the set of human body symmetries. The function g, which is H-equivariant, represents the unknown target function we aim to learn. In practice, g must respect the symmetries of the human body, meaning it must be H-equivariant. Thus, the assumption that g is H-equivariant is natural. Intuitively, Theorem 1 asserts that our Skeleton Mamba can effectively learn complex human motions, including those requiring precise coordination to follow both egocentric views and musical cues, while maintaining the equivariant properties associated with human body symmetries.

4.4. Training and Inference

Auxiliary Loss. As in state-of-the-art work in human motion generation [28, 69, 77], we employ position loss $\mathcal{L}_{pos}$ for accurate joint positioning and velocity loss $\mathcal{L}_{vel}$ for smooth motion dynamics. To reduce the foot sliding effects, we also use the contact loss $\mathcal{L}_{contact}$ as in [77]. The kinematic loss is expressed as follows:

$$\mathcal{L}_{kin} = \lambda_{pos}\mathcal{L}_{pos} + \lambda_{vel}\mathcal{L}_{vel} + \lambda_{contact}\mathcal{L}_{contact} \,. \tag{6}$$

Ego-Music Alignment Loss. We use $\mathcal{L}_{align}$ to align egocentric video and music at the temporal level. This loss ensures the high-dimensional features from both modalities are unified in a shared space, enabling the denoising model to generate human motion that is more coherent and contextually aligned with both inputs. Given the music embedding $\mathbf{z}_a \in \mathbb{R}^{T\times D_c}$ and egocentric vision embedding

$\mathbf{z}_v \in \mathbb{R}^{T \times D_c}$. This loss averages the symmetrical contrastive loss between vision and music embeddings:

$$\mathcal{L}_{align} = -\frac{1}{T}\sum_{i=1}^{T} \log \frac{\exp(\text{sim}(\mathbf{z}_a^i, \mathbf{z}_v^i)/\tau)}{\sum_{j=1}^{T} \exp(\text{sim}(\mathbf{z}_a^i, \mathbf{z}_v^j)/\tau)} , \quad (7)$$

where $\text{sim}(\mathbf{z}_a^i, \mathbf{z}_v^j)$ represents the cosine similarity between the music embedding at frame i and the vision embedding at frame j. The parameter τ is a temperature scalar that controls the sharpness of the similarity distribution.

We employ the diffusion loss $\mathcal{L}_{simple}$ as in [23]. The total training loss can be formulated as follows:

$$\mathcal{L}_{total} = \mathcal{L}_{simple} + \lambda_{kin}\mathcal{L}_{kin} + \lambda_{align}\mathcal{L}_{align} . \quad (8)$$

Head Guidance Sampling. To enhance the consistency of the head movements in the generated dance with egocentric images, we define a goal function $\mathcal{G}_{head}(\cdot)$ that guides the head to align closely with the head pose estimated from the egocentric images. This goal function consists of two components: a positional alignment term $\mathcal{G}_{pos}(\cdot)$ and a rotational alignment term $\mathcal{G}_{rot}(\cdot)$.

$$\begin{aligned} \mathcal{G}_{head}(\mathbf{x}) &= \gamma_{pos}\mathcal{G}_{pos}(\mathbf{x}) + \gamma_{rot}\mathcal{G}_{rot}(\mathbf{x}) , \\ \mathcal{G}_{pos}(\mathbf{x}) &= \frac{1}{T}\sum_{i=1}^{T} \|\mathbf{p}^i - \hat{\mathbf{p}}^i\|^2 , \\ \mathcal{G}_{rot}(\mathbf{x}) &= \frac{1}{T}\sum_{i=1}^{T} \| \log\left(\mathbf{R}^i \hat{\mathbf{R}}^{i\top}\right) \|_F^2 , \end{aligned} \quad (9)$$

where $\mathbf{p}^i$ and $\mathbf{R}^i$ represent the global head position and global head rotation matrix of the generated motion, respectively, and $\hat{\mathbf{p}}^i$ and $\hat{\mathbf{R}}^i$ are the corresponding estimated values calculated using the hybrid approach proposed in [43].

With goal function $\mathcal{G}_{head}(\cdot)$, we formulate the guided sampling problem as optimizing the probability of constraint satisfaction:

$$\begin{aligned} p(\mathbf{x}_0|\mathcal{O}=1,\mathbf{z}) &\propto p_\theta(\mathbf{x}_0|\mathbf{z})p(\mathcal{O}=1|\mathbf{x}_0,\mathbf{z}) \\ &\propto p_\theta(\mathbf{x}_0|\mathbf{z}) \cdot \exp(\mathcal{G}_{head}(\cdot)) , \end{aligned} \quad (10)$$

where $\mathcal{O}$ is an indicator to check if the generated dance motion $\mathbf{x}_m$ at denoising step m reaches the goal $\mathcal{G}_{head}(\cdot)$. Similar to [26], we use the first order Taylor expansion around $\mathbf{x}_m = \mu$ to estimate $p(\mathcal{O}=1|\mathbf{x}_m,\mathbf{z})$:

$$\log p(\mathcal{O}=1|\mathbf{x}_m,\mathbf{z}) \approx (\mathbf{x}_m - \mu)\xi + \mathcal{C} , \quad (11)$$

where $\mu = \mu_\theta(\mathbf{x}_m, m, \mathbf{z})$, $\mathcal{C}$ is a constant, and ξ is calculated as follows:

$$\begin{aligned} \xi &= \nabla_{\mathbf{x}_m} \log p(\mathcal{O}=1|\mathbf{x}_m,\mathbf{z})\big|_{\mathbf{x}_m=\mu} \\ &= \nabla_{\mathbf{x}_m} \mathcal{G}_{head}(\cdot)\big|_{\mathbf{x}_m=\mu} . \end{aligned} \quad (12)$$

Hence, we have the sampling process with a goal function:

$$p_\theta(\mathbf{x}_{m-1}|\mathbf{x}_m, \mathcal{O}=1, \mathbf{z}) = \mathcal{N}(\mathbf{x}_{m-1}; \mu + \lambda\Sigma\xi, \Sigma) , \quad (13)$$

here $\Sigma = \Sigma_\theta(\mathbf{x}_m, m, \mathbf{z})$ and λ is the scaling factor.

5. Experiments

Baselines. We compare our method EgoMusic Motion Network (EMM) with egocentric image-driven motion estimation works (PoseReg [91], Kinpoly [55], EgoEgo [43]) and music-driven motion generation works (FACT [45], Bailando [69], EDGE [77]). Since our task involves both egocentric video and music, for a fair comparison, we incorporate a music encoder, Jukebox [11], to process audio input for PoseReg [91], Kinpoly [55], EgoEgo [43]. For the music-driven works, we add visual features from the egocentric video using [22]. The implementation details of all baselines can be found in our Supplementary Material.

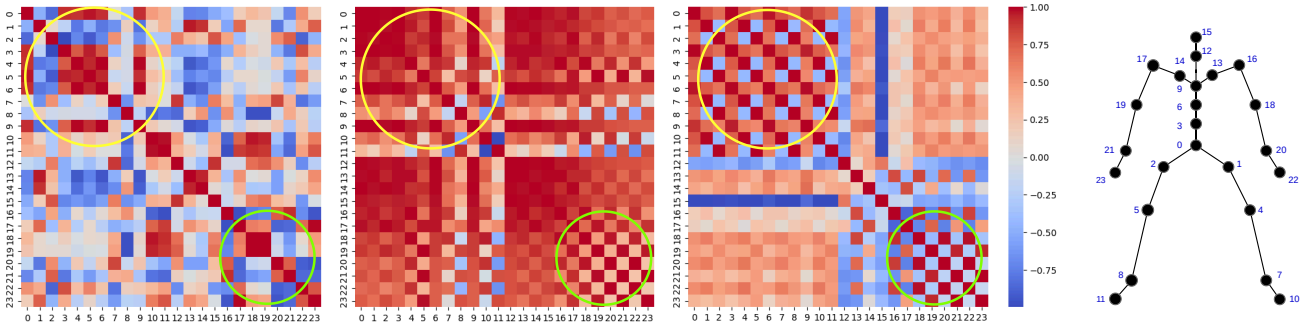
Metrics. Following [43], we evaluate our method using five standard metrics commonly used in human pose estimation task: *i)* $\mathbf{O}_{\text{head}}$, which measures the Frobenius norm between the predicted and actual head rotation matrices. *ii)* $\mathbf{T}_{\text{head}}$, calculated as the average Euclidean distance between the predicted and true head translation. *iii)* MPJPE, the Mean Per Joint Position Error, quantifies the average discrepancy in joint positions. *iv)* Accel refers to the difference in acceleration between predicted and ground truth joint positions. *v)* FS assesses foot skating, capturing unnatural foot movement. *vi)* To evaluate how well the generated dance motions align with the music and egocentric video, we propose a new metric called the Motion-Music-Vision (MMV) (please refer to our Supplementary Material).

5.1. Main Results

Baseline	$\mathbf{O}_{head}\downarrow$	$\mathbf{T}_{head}\downarrow$	MPJPE↓	Accel↓	FS↓	MMV↑
Pose-Reg [60]	1.78	423.56	351.37	37.14	98.79	0.182
Kinpoly [55]	1.16	392.67	338.74	16.27	25.81	0.197
EgoEgo [43]	0.74	373.67	152.02	14.23	22.13	0.218
FACT [45]	1.54	407.89	173.68	14.61	15.07	0.202
Bailando [69]	1.57	411.43	175.31	14.72	15.46	0.210
EDGE [77]	1.52	404.62	167.43	14.37	14.75	0.224
EMM (music only)	1.43	398.41	157.36	14.28	14.04	-
EMM (ego only)	0.61	355.16	186.49	16.02	13.45	-
EMM (ego + music)	**0.53**	**342.37**	**137.54**	**11.84**	**12.79**	**0.262**

Table 2. **Human motion estimation results.**

Table 2 presents the performance comparison between our method and other baselines. Table 2 shows that our EMM outperforms other baselines. EMM reduces head orientation error $\mathbf{O}_{head}$ by $0.21\,\text{rad}$ and head translation error $\mathbf{T}_{head}$ by $31.3\,\text{mm}$ compared to EgoEgo [43]. Furthermore, when compared to methods conditioned solely on music [45, 69, 77], our approach offers notable improvements in both accuracy in MPJPE and physical plausibility in FS metric. We also investigate the impact of each input modality, with results confirming that combining egocentric and music inputs enhances overall motion accuracy.



(a) Transformer (b) Vanilla Mamba (c) Skeleton Mamba (ours) (d) Human Skeleton

Figure 3. **Cosine similarity of joint embeddings.** We calculate the cosine similarity of joint embeddings of: a) Transformer [77], b) Vanilla Mamba [10], and c) Skelton Mamba (ours). d) Visualization of the human body with 24 joints for reference. We highlight similarities in arm joints (green circle) and leg joints (yellow circle). Our method shows a clear distinction between left and right limbs.

Baseline	$\mathbf{O}_{head}\downarrow$	$\mathbf{T}_{head}\downarrow$	MPJPE↓	Accel↓	FS↓	MMV↑
Pose-Reg [60]	1.21	642.47	377.56	30.36	56.18	0.165
Kinpoly [55]	0.78	354.19	251.27	17.84	25.31	0.187
EgoEgo [43]	0.67	347.23	234.58	16.76	20.15	0.203
FACT [45]	1.37	685.81	244.89	17.54	18.53	0.195
Bailando [69]	1.44	688.54	231.77	14.23	18.67	0.211
EDGE [77]	1.27	644.62	213.37	13.78	14.33	0.221
EMM (Ours)	**0.61**	**322.19**	**191.55**	**12.76**	**13.18**	**0.239**

Table 3. **Cross-dataset experiment results.**

5.2. Cross-dataset Results

To validate the generalization of our method, we conduct a cross-dataset experiment. We use the pretrained weights of all methods, originally trained on the EgoAIST++ dataset, to test on the EgoExo4D dataset [17]. EgoExo4D contains approximately two hours of dance sequences, along with egocentric videos and music captured in real-world. Table 3 shows that our method consistently outperforms other baselines in the cross-dataset experiment.

5.3. Skeleton Mamba Analysis

Can Skeleton Mamba learn human skeleton? To answer this, we compute the cosine similarity matrices of joint embeddings for Transformer [77], Vanilla Mamba [10], and our Skeleton Mamba. Fig. 3 shows that Skeleton Mamba captures the human body structure more clearly. We highlight the similarity between joints in the arms (green circles) and the legs (yellow circles) in Fig. 3. In particular, the joints of the right arm (join 17, 19, 21, and 23 in Fig. 3d) exhibit strong similarity to each other. By contrast, joint 20, which belongs to the left arm, shows low similarity with the joints in the right arm, indicating effective separation between limbs. In comparison, Transformer [77] and Vanilla Mamba [10] models display less distinct differentiation.

Scan strategy analysis. Table 4 shows the impact of different scanning schemes on learning human motion. When applying the scanning to the spatial domain, unidirectional scanning [10] and bidirectional scanning [96] show reasonable results but are lower than our Skeleton Mamba. Table 4 also shows that our method, which uses bidirectional scanning (i.e., ForwardSSD and BackwardSDD in Equation 5) for temporal processing, outperforms the unidirectional approach. This experiment and Theorem 1 confirm that by learning in both temporal and spatial domains, our Skeleton Mamba can effectively handle the geometry of human motion that aligns with both egocentric and music input.

Dimension	Scan Type	$\mathbf{O}_{head}\downarrow$	$\mathbf{T}_{head}\downarrow$	MPJPE↓	Accel↓	FS↓	MMV↑
Spatial	Unidirectional [10]	0.63	386.12	192.38	16.21	14.65	0.245
	Bidirectional [96]	0.58	357.81	160.12	14.02	14.12	0.255
Temporal	Unidirectional [10]	0.61	371.62	181.43	15.29	15.47	0.251
Temporal & Spatial	Skeleton Mamba	**0.53**	**342.37**	**137.54**	**11.84**	**12.79**	**0.262**

Table 4. **Scanning strategy analysis.**

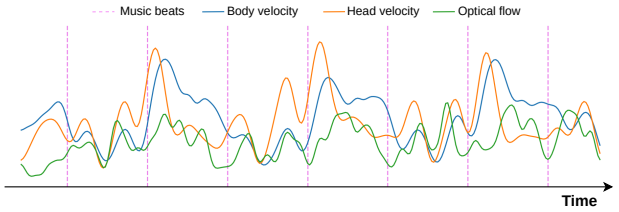


Figure 4. **Motion, music, and egocentric view correlation.**

5.4. Ablation Study

Motion, music, and egocentric correlation. We plot the kinematic velocity, music beats, and optical flow extracted from the egocentric video to visualize the correlation between these three. The music beats are calculated using the beat extracting algorithm [58]. The motion beats are extracted as the local extrema of the kinematic velocity. The optical flow is extracted from the egocentric video using RAFT [73]. Fig. 4 shows that the generated dance aligns well with the music beat and optical flow. The results show that our model effectively synchronizes both body and head movements with the respective audio and visual cues, generating well-coordinated motion from multimodal inputs.

Ablation	$\mathbf{O}_{head}\downarrow$	$\mathbf{T}_{head}\downarrow$	MPJPE↓	Accel↓	FS↓	MMV↑
EMM	0.53	342.37	137.54	11.84	12.79	0.262
EMM w.o. $\mathcal{G}_{head}$	0.76	374.86	146.38	12.71	13.34	0.254
EMM w.o. $\mathcal{L}_{align}$	0.56	350.06	142.47	12.16	13.09	0.242

Table 5. **Loss analysis**.

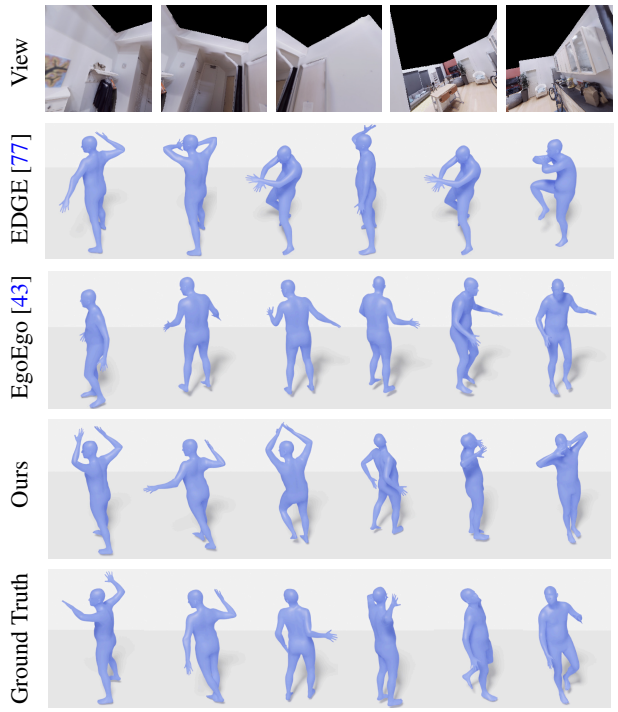


Figure 5. **Qualitative comparison.** Our approach produces well-aligned human motion with ego view and music.

Loss analysis. Table 5 shows that each loss function plays an essential role in our network. In particular, $\mathcal{L}_{align}$ positively impacts all metrics, especially in terms of motion-music-vision synchronization. Additionally, the head guidance loss $\mathcal{G}_{head}$, applied during the sampling process, substantially increases the accuracy of head movements, ensuring better consistency with the egocentric view.

Quality results. Fig. 5 shows qualitative comparisons between EDGE [77], EgoEgo [43] and our method. EgoEgo [43] has difficulty generating dance motions that follow the choreography. EDGE [77] generates dance movements in sync with the music, but its head movements do not correspond to the egocentric input. In contrast, our approach demonstrates a more cohesive generation of both dance and head movements.

5.5. Skeleton Mamba on Human Motion Tasks

To further evaluate the effectiveness of our Skeleton Mamba in modeling human motion, we conduct experiments on text-to-human motion generation and skeleton-based action recognition tasks. Implementation details of both tasks can be found in our Supplementary Material.

Method	R Prec.↑	MM Dist↓	Div→
Ground Truth	0.797	2.974	9.503
MDM [74]	0.611	5.566	9.559
MLD [8]	0.772	3.196	9.724
MMamba [96]	0.790	3.060	9.871
Ours	**0.795**	**2.983**	**9.484**

Table 6. **Text2motion results.**

Method	NTU60-XS	Kinetics
MS-G3D [53]	91.5	38.0
PoseConv3D [14]	93.1	47.7
MotionBERT [100]	93.0	-
DSTA-Net [68]	91.5	-
Ours	**94.4**	**52.4**

Table 7. **Action rec. results.**

Text-to-motion generation. We evaluate our Skeleton Mamba on HumanML3D [19] dataset. We compare with recent work MDM [74], MLD [8] and [96] using metrics in [19]. Table 6 shows that our method outperforms other baselines in the text-to-human motion generation task.

Human action recognition. We benchmark on two datasets: Kinetics400 [31] and NTU RGB+D 60 [49, 67]. We compare with recent methods, including MS-G3D [53], PoseConv3D [14], MotionBERT [100] and DSTA-Net [68]. Top-1 classification accuracy is used as the metric. Table 7 shows that our method clearly outperforms the other action recognition baselines. Tables 6 and 7 demonstrate that while our Skeleton Mamban is designed for human motion modeling, it has the potential to generalize effectively to various setups, including generative and recognition tasks.

6. Discussion

Broader Impact. We believe our work represents a significant step toward understanding human motion dynamics and potentially has a profound impact on different fields such as VR/AR, metaverse, or film animation. For instance, in VR dance games [7, 34, 66], current systems rely on egocentric cameras and additional motion tracking devices such as hand VR motion controllers. By fusing music cues with egocentric data, however, full-body motion can be accurately estimated, eliminating the need for additional sensors. Moreover, our Skeleton Mamba has the potential in other tasks such as human motion synthesis [74], human action recognition [76, 88, 90], gesture analysis [51], and human-object interaction [32].

Limitations. Although our method achieves encouraging results, it still presents certain limitations. First, while our model effectively generates smooth motion sequences, it is not fully optimized for producing very long motion sequences due to the reliance on a simple bidirectional scan in the temporal dimension. Second, when the egocentric video and the music input are not appropriately paired, our model may fail to generate coherent and synchronized motion.

Conclusion. We introduce a new task and method to estimate human dance motion from egocentric and music. Our network with the core Skeleton Mamba effectively estimates motion that aligns with both visual and musical cues. We further contribute EgoAIST++ dataset, which provides egocentric, music, and dance groundtruth. Intensive experiments show that our method significantly outperforms existing state-of-the-art approaches, and our Skeleton Mamba has the potential in human motion understanding tasks.

References

[1] Hiroyasu Akada, Jian Wang, Soshi Shimada, Masaki Takahashi, Christian Theobalt, and Vladislav Golyanik. Unrealego: A new dataset for robust egocentric 3d human motion capture. In *ECCV*, 2022.

[2] Ho Yin Au, Jie Chen, Junkun Jiang, and Yike Guo. Choreograph: Music-conditioned automatic dance choreography over a style and tempo consistent dynamic graph. In *ACMMM*, 2022.

[3] Ali Behrouz and Farnoosh Hashemi. Graph mamba: Towards learning on graphs with state space models. In *KDD*, 2024.

[4] Asish Bera, Mita Nasipuri, Ondrej Krejcar, and Debotosh Bhattacharjee. Fine-grained sports, yoga, and dance postures recognition: A benchmark analysis. *TIM*, 2023.

[5] Steven Brown and Lawrence M Parsons. The neuroscience of dance. *Scientific American*, 2008.

[6] Caroline Chan, Shiry Ginosar, Tinghui Zhou, and Alexei A Efros. Everybody dance now. In *ICCV*, 2019.

[7] Jacky CP Chan, Howard Leung, Jeff KT Tang, and Taku Komura. A virtual reality dance training system using motion capture technology. *IEEE TLT*, 2010.

[8] Xin Chen, Biao Jiang, Wen Liu, Zilong Huang, Bin Fu, Tao Chen, and Gang Yu. Executing your commands via motion diffusion in latent space. In *CVPR*, 2023.

[9] Rishabh Dabral, Muhammad Hamza Mughal, Vladislav Golyanik, and Christian Theobalt. Mofusion: A framework for denoising-diffusion-based motion synthesis. In *CVPR*, 2023.

[10] Tri Dao and Albert Gu. Transformers are SSMs: Generalized models and efficient algorithms through structured state space duality. In *ICML*, 2024.

[11] Prafulla Dhariwal, Heewoo Jun, Christine Payne, Jong Wook Kim, Alec Radford, and Ilya Sutskever. Jukebox: A generative model for music. *arXiv:2005.00341*, 2020.

[12] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *NeurIPS*, 2021.

[13] Gisela Miranda Difini, Marcio Garcia Martins, and Jorge Luis Victória Barbosa. Human pose estimation for training assistance: a systematic literature review. In *BSMW*, 2021.

[14] Haodong Duan, Yue Zhao, Kai Chen, Dahua Lin, and Bo Dai. Revisiting skeleton-based action recognition. In *CVPR*, 2022.

[15] Di Fan, Lili Wan, Wanru Xu, and Shenghui Wang. A bidirectional attention guided cross-modal network for music based dance generation. *Computers and Electrical Engineering*, 2022.

[16] Rukun Fan, Songhua Xu, and Weidong Geng. Example-based automatic music-driven conventional dance motion synthesis. *IEEE transactions on visualization and computer graphics*, 2011.

[17] Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, et al. Ego-exo4d: Understanding skilled human activity from first-and third-person perspectives. In *CVPR*, 2024.

[18] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv:2312.00752*, 2023.

[19] Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating diverse and natural 3d human motions from text. In *CVPR*, 2022.

[20] Hong Guo, ShanChen Zou, YiLin Xu, Han Yang, Jian Wang, HongXin Zhang, and Wei Chen. Dancevis: toward better understanding of online cheer and dance training. *Journal of Visualization*, 2022.

[21] Judith Lynne Hanna. *To dance is human: A theory of nonverbal communication*. University of Chicago Press, 1987.

[22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[23] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NeurIPS*, 2020.

[24] Vincent Tao Hu, Stefan Andreas Baumann, Ming Gui, Olga Grebenkova, Pingchuan Ma, Johannes Fischer, and Björn Ommer. Zigma: A dit-style zigzag mamba diffusion model. In *ECCV*, 2024.

[25] Xiaodan Hu and Narendra Ahuja. Unsupervised 3d pose estimation for hierarchical dance video recognition. In *ICCV*, 2021.

[26] Siyuan Huang, Zan Wang, Puhao Li, Baoxiong Jia, Tengyu Liu, Yixin Zhu, Wei Liang, and Song-Chun Zhu. Diffusion-based generation, optimization, and planning in 3d scenes. In *CVPR*, 2023.

[27] Yuhang Huang, Junjie Zhang, Shuyan Liu, Qian Bao, Dan Zeng, Zhineng Chen, and Wu Liu. Genre-conditioned long-term 3d dance generation driven by music. In *ICASSP*, 2022.

[28] Zikai Huang, Xuemiao Xu, Cheng Xu, Huaidong Zhang, Chenxi Zheng, Jing Qin, and Shengfeng He. Beat-it: Beat-synchronized multi-condition 3d dance generation. *arXiv:2407.07554*, 2024.

[29] Hao Jiang and Kristen Grauman. Seeing invisible poses: Estimating 3d body pose from egocentric video. In *CVPR*, 2017.

[30] Hao Jiang and Vamsi Krishna Ithapu. Egocentric pose estimation from human vision span. In *ICCV*, 2021.

[31] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv:1705.06950*, 2017.

[32] Bumsoo Kim, Junhyun Lee, Jaewoo Kang, Eun-Sol Kim, and Hyunwoo J Kim. Hotr: End-to-end human-object interaction detection with transformers. In *CVPR*, 2021.

[33] Hyeonyeong Kim and Hwansoo Lee. Performing arts metaverse: The effect of perceived distance and subjective experience. *Computers in Human Behavior*, 2023.

[34] Markus Laattala, Roosa Piitulainen, Nadia M Ady, Monica Tamariz, and Perttu Hämäläinen. Wave: Anticipatory movement visualization for vr dancing. In *CHFCS*, 2024.

[35] Kit Yung Lam, Liang Yang, Ahmad Alhilal, Lik-Hang Lee, Gareth Tyson, and Pan Hui. Human-avatar interaction in metaverse: Framework for full-body interaction. In *ACM and ICMA*, 2022.

[36] Kimerer LaMothe. The dancing species: how moving together in time helps make us human. *Aeon, June*, 2019.

[37] Nhat Le, Khoa Do, Xuan Bui, Tuong Do, Erman Tjiputra, Quang D Tran, and Anh Nguyen. Scalable group choreography via variational phase manifold learning. In *ECCV*, 2024.

[38] Nhat Le, Tuong Do, Khoa Do, Hien Nguyen, Erman Tjiputra, Quang D Tran, and Anh Nguyen. Controllable group choreography using contrastive diffusion. *ACM Transactions on Graphics (TOG)*, 2023.

[39] Nhat Le, Thang Pham, Tuong Do, Erman Tjiputra, Quang D Tran, and Anh Nguyen. Music-driven group choreography. In *CVPR*, 2023.

[40] Byung-Kwan Lee, Chae Won Kim, Beomchan Park, and Yong Man Ro. Meteor: Mamba-based traversal of rationale for large language and vision models. *NeuRIPS*, 2024.

[41] Minho Lee, Kyogu Lee, and Jaeheung Park. Music similarity-based approach to generating dance motion sequence. *Multimedia tools and applications*, 2013.

[42] Buyu Li, Yongchi Zhao, Shi Zhelun, and Lu Sheng. Danceformer: Music conditioned 3d dance generation with parametric motion transformer. In *AAAI*, 2022.

[43] Jiaman Li, Karen Liu, and Jiajun Wu. Ego-body pose estimation via ego-head pose estimation. In *CVPR*, 2023.

[44] Kunchang Li, Xinhao Li, Yi Wang, Yinan He, Yali Wang, Limin Wang, and Yu Qiao. Videomamba: State space model for efficient video understanding. In *ECCV*, 2025.

[45] Ruilong Li, Shan Yang, David A Ross, and Angjoo Kanazawa. Ai choreographer: Music conditioned 3d dance generation with aist++. In *ICCV*, 2021.

[46] Shufan Li, Harkanwar Singh, and Aditya Grover. Mamba-nd: Selective state space modeling for multi-dimensional data. In *ECCV*, 2025.

[47] Tianyi Li, Chi Zhang, Wei Su, and Yuehu Liu. Egoformer: Transformer-based motion context learning for ego-pose estimation. In *SMC*, 2023.

[48] Dingkang Liang, Xin Zhou, Wei Xu, Xingkui Zhu, Zhikang Zou, Xiaoqing Ye, Xiao Tan, and Xiang Bai. Pointmamba: A simple state space model for point cloud analysis. In *NeurIPS*, 2024.

[49] Jun Liu, Amir Shahroudy, Mauricio Perez, Gang Wang, Ling-Yu Duan, and Alex C Kot. Ntu rgb+ d 120: A large-scale benchmark for 3d human activity understanding. *TPAMI*, 2019.

[50] Jiuming Liu, Ruiji Yu, Yian Wang, Yu Zheng, Tianchen Deng, Weicai Ye, and Hesheng Wang. Point mamba: A novel point cloud backbone based on state space model with octree-based ordering strategy. *arXiv:2403.06467*, 2024.

[51] Xin Liu, Henglin Shi, Haoyu Chen, Zitong Yu, Xiaobai Li, and Guoying Zhao. imigue: An identity-free video dataset for micro-gesture understanding and emotion analysis. In *CVPR*, 2021.

[52] Yuxuan Liu, Jianxin Yang, Xiao Gu, Yijun Chen, Yao Guo, and Guang-Zhong Yang. Egofish3d: Egocentric 3d pose estimation from a fisheye camera via self-supervised learning. *IEEE Transactions on Multimedia*, 2023.

[53] Ziyu Liu, Hongwen Zhang, Zhenghao Chen, Zhiyong Wang, and Wanli Ouyang. Disentangling and unifying graph convolutions for skeleton-based action recognition. In *CVPR*, 2020.

[54] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. In *Seminal Graphics Papers: Pushing the Boundaries*. 2023.

[55] Zhengyi Luo, Ryo Hachiuma, Ye Yuan, and Kris Kitani. Dynamics-regulated kinematic policy for egocentric pose estimation. *NeurIPS*, 2021.

[56] Hitoshi Matsuyama, Shunsuke Aoki, Takuro Yonezawa, Kei Hiroi, Katsuhiko Kaji, and Nobuo Kawaguchi. Deep learning for ballroom dance recognition: A temporal and trajectory-aware classification model with three-dimensional pose estimation and wearable sensing. *IEEE Sensors Journal*, 2021.

[57] Hitoshi Matsuyama, Kei Hiroi, Katsuhiko Kaji, Takuro Yonezawa, and Nobuo Kawaguchi. Hybrid activity recognition for ballroom dance exercise using video and wearable sensor. In *ICIEV and icIVPR*, 2019.

[58] Brian McFee, Colin Raffel, Dawen Liang, Daniel PW Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. librosa: Audio and music signal analysis in python. In *SciPy*, 2015.

[59] Christen Millerdurai, Hiroyasu Akada, Jian Wang, Diogo Luvizon, Christian Theobalt, and Vladislav Golyanik. Eventego3d: 3d human motion capture from egocentric event streams. In *CVPR*, 2024.

[60] Evonne Ng, Donglai Xiang, Hanbyul Joo, and Kristen Grauman. You2me: Inferring body pose in egocentric video via first and second person interactions. In *CVPR*, 2020.

[61] Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *AAAI*, 2018.

[62] Chiara Plizzari, Marco Cannici, and Matteo Matteucci. Spatial temporal transformer network for skeleton-based action recognition. In *ICPR*, 2021.

[63] Ziyun Qian, Zeyu Xiao, Zhenyi Wu, Dingkang Yang, Mingcheng Li, Shunli Wang, Shuaibing Wang, Dongliang Kou, and Lihua Zhang. Smcd: High realism motion style transfer via mamba-based diffusion. *arXiv:2405.02844*, 2024.

[64] Li Qianwen. Application of motion capture technology based on wearable motion sensor devices in dance body motion recognition. *Measurement Sensors*, 2024.

[65] Challapalli Jhansi Rani and Nagaraju Devarakonda. An effectual classical dance pose estimation and classification system employing convolution neural network–long short-term memory (cnn-lstm) network for video sequences. *Microprocessors and Microsystems*, 2022.

[66] Bhuvaneswari Sarupuri, Richard Kulpa, Andreas Aristidou, and Franck Multon. Dancing in virtual reality as an inclusive platform for social and physical fitness activities: A survey. *The Visual Computer*, 2024.

[67] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In *CVPR*, 2016.

[68] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. Decoupled spatial-temporal attention network for skeleton-based action-gesture recognition. In *ACCV*, 2020.

[69] Li Siyao, Weijiang Yu, Tianpei Gu, Chunze Lin, Quan Wang, Chen Qian, Chen Change Loy, and Ziwei Liu. Bailando: 3d dance generation by actor-critic gpt with choreographic memory. In *CVPR*, 2022.

[70] Julian Straub, Thomas Whelan, Lingni Ma, Yufan Chen, Erik Wijmans, Simon Green, Jakob J Engel, Raul Mur-Artal, Carl Ren, Shobhit Verma, et al. The replica dataset: A digital replica of indoor spaces. *arXiv:1906.05797*, 2019.

[71] Guofei Sun, Yongkang Wong, Zhiyong Cheng, Mohan S Kankanhalli, Weidong Geng, and Xiangdong Li. Deepdance: music-to-dance motion choreography with adversarial learning. *IEEE Transactions on Multimedia*, 2020.

[72] Andrew Szot, Alexander Clegg, Eric Undersander, Erik Wijmans, Yili Zhao, John Turner, Noah Maestre, Mustafa Mukadam, Devendra Singh Chaplot, Oleksandr Maksymets, et al. Habitat 2.0: Training home assistants to rearrange their habitat. *NeurIPS*, 2021.

[73] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *ECCV*, 2020.

[74] Guy Tevet, Sigal Raab, Brian Gordon, Yoni Shafir, Daniel Cohen-or, and Amit Haim Bermano. Human motion diffusion model. In *ICLR*, 2023.

[75] Denis Tome, Patrick Peluse, Lourdes Agapito, and Hernan Badino. xr-egopose: Egocentric 3d human pose from an hmd camera. In *ICCV*, 2019.

[76] Shashank Tripathi, Lea Müller, Chun-Hao P Huang, Omid Taheri, Michael J Black, and Dimitrios Tzionas. 3d human pose estimation via intuitive physics. In *CVPR*, 2023.

[77] Jonathan Tseng, Rodrigo Castellon, and Karen Liu. Edge: Editable dance generation from music. In *CVPR*, 2023.

[78] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017.

[79] Chloe Wang, Oleksii Tsepa, Jun Ma, and Bo Wang. Graph-mamba: Towards long-range graph sequence modeling with selective state spaces. *arXiv:2402.00789*, 2024.

[80] Jian Wang, Zhe Cao, Diogo Luvizon, Lingjie Liu, Kripasindhu Sarkar, Danhang Tang, Thabo Beeler, and Christian Theobalt. Egocentric whole-body motion capture with fisheyevit and diffusion-based motion refinement. In *CVPR*, 2024.

[81] Jian Wang, Lingjie Liu, Weipeng Xu, Kripasindhu Sarkar, and Christian Theobalt. Estimating egocentric 3d human pose in global space. In *ICCV*, 2021.

[82] Jian Wang, Diogo Luvizon, Weipeng Xu, Lingjie Liu, Kripasindhu Sarkar, and Christian Theobalt. Scene-aware egocentric 3d human pose estimation. *CVPR*, 2023.

[83] Jiashun Wang, Huazhe Xu, Jingwei Xu, Sifei Liu, and Xiaolong Wang. Synthesizing long-term 3d human motion and interaction in 3d scenes. In *CVPR*, 2021.

[84] Tan Wang, Linjie Li, Kevin Lin, Yuanhao Zhai, Chung-Ching Lin, Zhengyuan Yang, Hanwang Zhang, Zicheng Liu, and Lijuan Wang. Disco: Disentangled control for realistic human dance generation. In *CVPR*, 2024.

[85] Xinghan Wang, Zixi Kang, and Yadong Mu. Text-controlled motion mamba: Text-instructed temporal grounding of human motion. *arXiv:2404.11375*, 2024.

[86] Ziyang Wang, Jian-Qing Zheng, Yichi Zhang, Ge Cui, and Lei Li. Mamba-unet: Unet-like pure visual mamba for medical image segmentation. *arXiv:2402.05079*, 2024.

[87] Weipeng Xu, Avishek Chatterjee, Michael Zollhoefer, Helge Rhodin, Pascal Fua, Hans-Peter Seidel, and Christian Theobalt. Mo^2Cap^2 : Real-time mobile 3d motion capture with a cap-mounted fisheye camera. *IEEE Transactions on Visualization and Computer Graphics*, 2019.

[88] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. Vitpose: Simple vision transformer baselines for human pose estimation. *NeurIPS*, 2022.

[89] Jing Nathan Yan, Jiatao Gu, and Alexander M Rush. Diffusion models without attention. In *CVPR*, 2024.

[90] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *AAAI*, 2018.

[91] Ye Yuan and Kris Kitani. Ego-pose estimation and forecasting as real-time pd control. In *ICCV*, 2019.

[92] Canyu Zhang, Youbao Tang, Ning Zhang, Ruei-Sung Lin, Mei Han, Jing Xiao, and Song Wang. Bidirectional autoregessive diffusion model for dance generation. In *CVPR*, 2024.

[93] Guowen Zhang, Lue Fan, Chenhang He, Zhen Lei, ZHAO-XIANG ZHANG, and Lei Zhang. Voxel mamba: Group-free state space models for point cloud based 3d object detection. *NeuRIPS*, 2024.

[94] Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. Motiondiffuse: Text-driven human motion generation with diffusion model. *arXiv:2208.15001*, 2022.

[95] Zeyu Zhang, Akide Liu, Qi Chen, Feng Chen, Ian Reid, Richard Hartley, Bohan Zhuang, and Hao Tang. Infinimotion: Mamba boosts memory in transformer for arbitrary long motion generation. *arXiv:2407.10061*, 2024.

[96] Zeyu Zhang, Akide Liu, Ian Reid, Richard Hartley, Bohan Zhuang, and Hao Tang. Motion mamba: Efficient and long sequence motion generation. In *ECCV*, 2024.

[97] Amy Zhao, Chengcheng Tang, Lezi Wang, Yijing Li, Mihika Dave, Lingling Tao, Christopher D. Twigg, and Robert Y. Wang. Egobody3m: Egocentric body tracking on a vr headset using a diverse dataset. In *ECCV*, 2024.

[98] Jiaojiao Zhao and Cees GM Snoek. Dance with flow: Two-in-one stream action detection. In *CVPR*, 2019.

[99] Lianghui Zhu, Bencheng Liao, Qian Zhang, Xinlong Wang, Wenyu Liu, and Xinggang Wang. Vision mamba: Efficient visual representation learning with bidirectional state space model. *arXiv:2401.09417*, 2024.

[100] Wentao Zhu, Xiaoxuan Ma, Zhaoyang Liu, Libin Liu, Wayne Wu, and Yizhou Wang. Motionbert: A unified perspective on learning human motion representations. In *ICCV*, 2023.

[101] Wenlin Zhuang, Congyi Wang, Siyu Xia, Jinxiang Chai, and Yangang Wang. Music2dance: Music-driven dance generation using wavenet. *arXiv:2002.03761*, 2020.