AI Choreographer: Music Conditioned 3D Dance Generation with AIST++

Ruilong Li[*1] Shan Yang[*2] David A. Ross[2] Angjoo Kanazawa[2,3]
[1]University of Southern California [2]Google Research [3]University of California, Berkeley

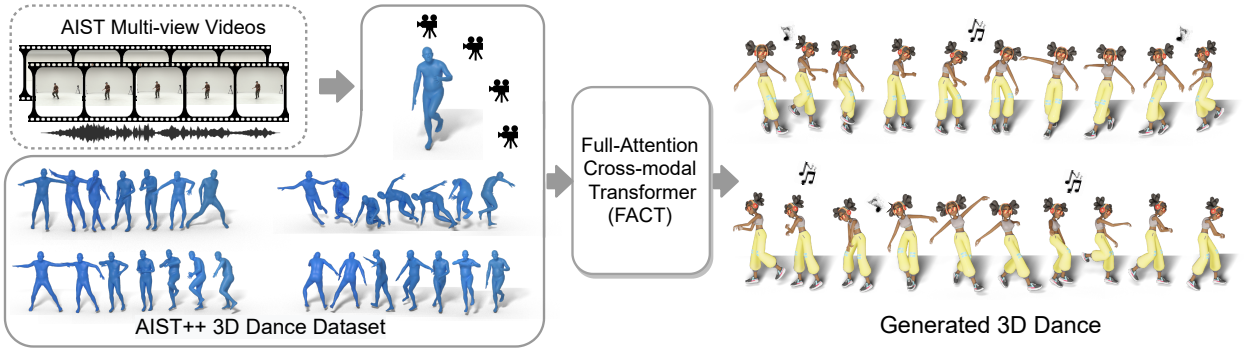

Figure 1: **AI Choreographer.** We present a new 3D dance dataset, AIST++, which contains 5.2 hours of 3D motion reconstructed from real dancers paired with music (left) and a novel Full-Attention Cross-modal Transformer (FACT) network that can generate realistic 3D dance motion with global translation conditioned on music (right). We output our 3D motion in representations that allow for instant motion retargeting to a novel character. Here we use a character from Mixamo [1]

Abstract

We present AIST++, a new multi-modal dataset of 3D dance motion and music, along with FACT, a Full-Attention Cross-modal Transformer network for generating 3D dance motion conditioned on music. The proposed AIST++ dataset contains 5.2 hours of 3D dance motion in 1408 sequences, covering 10 dance genres with multi-view videos with known camera poses—the largest dataset of this kind to our knowledge. We show that naively applying sequence models such as transformers to this dataset for the task of music conditioned 3D motion generation does not produce satisfactory 3D motion that is well correlated with the input music. We overcome these shortcomings by introducing key changes in its architecture design and supervision: FACT model involves a deep cross-modal transformer block with full-attention that is trained to predict N future motions. We empirically show that these changes are key factors in generating long sequences of realistic dance motion that are well-attuned to the input music. We conduct extensive experiments on AIST++ with user studies, where our method outperforms recent state-of-the-art methods both qualitatively and quantitatively. The code and the dataset can be found at: https://google.github.io/aichoreographer.

1. Introduction

The ability to dance by composing movement patterns that align to musical beats is a fundamental aspect of human behavior. Dancing is an universal language found in all cultures [50], and today, many people express themselves through dance on contemporary online media platforms. The most watched videos on YouTube are dance-centric music videos such as "Baby Shark Dance", and "Gangnam Style" [75], making dance a more and more powerful tool to spread messages across the internet. However, dancing is a form of art that requires practice—even for humans, professional training is required to equip a dancer with a rich repertoire of dance motions to create an expressive choreography. Computationally, this is even more challenging as the task requires the ability to generate a continuous motion with high kinematic complexity that captures the non-linear relationship with the accompanying music.

In this work, we address these challenges by presenting a novel Full Attention Cross-modal Transformer (FACT) network, which can robustly generate realistic 3D dance motion from music, along with a large-scale multi-modal 3D dance motion dataset, AIST++, to train such a model. Specifically, given a piece of music and a short (2 seconds) seed motion, our model is able to generate a long sequence of realistic 3D dance motions. Our model effectively learns the music-motion correlation and can generate dance se-

[*] equal contribution. Work performed while Ruilong was an intern at Google.

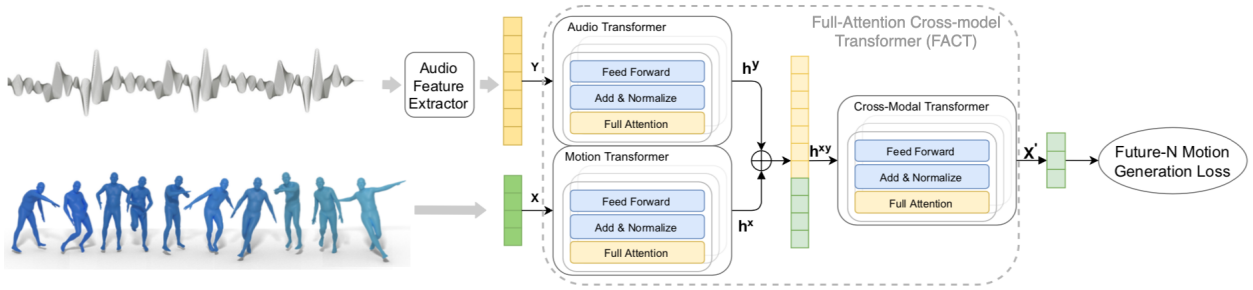

Figure 2: **Cross-Modal Music Conditioned 3D Motion Generation Overview.** Our proposed a Full-Attention Cross-modal Transformer (FACT) network (details in Figure 3) takes in a music piece and a 2-second sequence of seed motion, then auto-regressively generates long-range future motions that correlates with the input music.

quences that varies for different input music. We represent dance as a 3D motion sequence that consists of joint rotation and global translation, which enables easy transfer of our output for applications such as motion retargeting as shown in Figure 1.

In order to generate 3D dance motion from music, we propose a novel Full Attention Cross-modal Transformer (FACT) model, which employs an audio transformer and seed motion transformer to encode the inputs, which are then fused by a cross-modal transformer that models the distribution between audio and motion. This model is trained to predict N future motion sequences and at test time is applied in an auto-regressive manner to generate continuous motion. The success of our model relies on three key design choices: 1) the use of full-attention in an auto-regressive model, 2) future-N supervision, and 3) early fusion of two modalities. The combination of these choices is critical for training a model that can generate a long realistic dance motion that is attuned to the music. Although prior work has explored using transformers for motion generation [3], we find that naively applying transformers to the 3D dance generation problem without these key choices does not lead to a very effective model.

In particular, we notice that because the context window in the motion domain is significantly smaller than that of language models, it is possible to apply full-attention transformers in an auto-regressive manner, which leads to a more powerful model. It is also critical that the full-attention transformer is trained to predict N possible future motions instead of one. These two design choices are key for preventing 3D motion from freezing or drifting after several auto-regressive steps as reported in prior works on 3D motion generation [4, 3]. Our model is trained to predict 20 future frames, but it is able to produce realistic 3D dance motion for over 1200 frames at test time. We also show that fusing the two modalities early, resulting in a deep cross-modal transformer, is important for training a model that generates different dance sequences for different music.

In order to train the proposed model, we also address the problem of data. While there are a few motion capture datasets of dancers dancing to music, collecting mocap data requires heavily instrumented environments making these datasets severely limited in the number of available dance sequences, dancer and music diversity. In this work, we propose a new dataset called AIST++, which we build from the existing multi-view dance video database called AIST [82]. We use the multi-view videos to recover reliable 3D motion from this data. We will release code and this dataset for research purposes, where AIST++ can be a new benchmark for the task of 3D dance generation conditioned on music.

In summary, our contributions are as follows:

- We propose Full Attention Cross-Modal Transformer model, FACT, which can generate a long sequence of realistic 3D dance motion that is well correlated with the input music.
- We introduce AIST++ dataset containing 5.2 hours of 3D dance motions accompanied with music and multi-view images, which to our knowledge is the largest dataset of such kind.
- We provide extensive evaluations validating our design choices and show that they are critical for high quality, multi-modal, long motion sequence generation.

2. Related Work

3D Human Motion Synthesis The problem of generating realistic and controllable 3D human motion sequences has long been studied. Earlier works employ statistical models such as kernel-based probability distribution [64, 10, 25, 11] to synthesize motion, but abstract away motion details. Motion graphs [53, 7, 47] address this problem by generating motions in a non-parametric manner. Motion graph is a directed graph constructed on a corpus of motion capture data, where each node is a pose and the edges represent the transition between poses. Motion is generated by a random walk on this graph. A challenge in motion graph is in generating plausible transition that some approaches address via parameterizing the transition [30]. With the development in deep learning, many approaches explore the applicability of neural networks to generate 3D motion by training on a large-scale motion capture dataset, where network architectures such as CNNs [35, 34], GANs [31], RBMs [80],

RNNs [24, 4, 40, 27, 16, 18, 88, 12, 87] and Transformers [3, 9] have been explored. Auto-regressive models like RNNs and vanilla Transformers are capable of generating unbounded motion in theory, but in practice suffer from regression to the mean where motion "freezes" after several iterations, or drift to unnatural motions [4, 3]. Some works [8, 56, 49] propose to ease this problem by periodically using the network's own outputs as inputs during training. Phase-functioned neural networks and it's variations [94, 33, 73, 74] address this issue via conditioning the network weights on phase, however, they do not scale well to represent a wide variety of motion.

Audio To Human Motion Generation Audio to motion generation has been studied in 2D pose context either in optimization based approach [81], or learning based approaches [52, 72, 51, 67, 68, 21] where 2D pose skeletons are generated from a conditioning audio. Training data for 2D pose and audio is abundant thanks to the high reliability of 2D pose detectors [14]. However, predicting motion in 2D is limited in its expressiveness and potential for downstream applications. For 3D dance generation, earlier approaches explore matching existing 3D motion to music [71] using motion graph based approach [20]. More recent approach employ LSTMs [5, 79, 90, 97, 42], GANs [51, 78, 28], transformer encoder with RNN decoder [36] or convolutional [2, 92] sequence-to-sequence models. Concurrent to our work, Chen *et al*. [15] proposed a method that is based on motion graphs with learned embedding space. Many prior works [72, 68, 42, 28, 92] solve this problem by predicting future motion deterministically from audio without seed motion. When the same audio has multiple corresponding motions, which often occurs in dance data, these methods collapse to predicting a mean pose. In contrast, we formulate the problem with seed motion as in [55, 96], which allows generation of multiple motion from the same audio even with a deterministic model.

Closest to our work is that of Li *et al*. [55], which also employ transformer based architecture but only on audio and motion. Furthermore, their approach discretize the output joint space in order to account for multi-modality, which generates unrealistic motion. In this work we introduce a novel full-attention based cross-modal transformer (FACT model) for audio and motion, which can not only preserve the correlation between music and 3D motion better, but also generate more realistic long 3D human motion with global translation. One of the biggest bottleneck in 3D dance generation approaches is that of data. Recent work of Li *et al*. [55] reconstruct 3D motion from dance videos on the Internet, however the data is not public. Further, using 3D motion reconstructed from monocular videos may not be reliable and lack accurate global 3D translation information. In this work we also reconstruct the 3D motion from 2D dance video, but from multi-view video sequences, which addresses these issues. While there are many large scale 3D motion capture datasets [39, 59, 1, 37], mocap dataset of 3D dance is quite limited as it requires heavy instrumentation and expert dancers for capture. As such, many of these previous works operate on either small-scale or private motion capture datasets [79, 5, 96]. We compare our proposed dataset with these public datasets in Table 1.

Cross-Modal Sequence-to-Sequence Generation Beyond of the scope of human motion generation, our work is closely related to the research of using neural network on cross-modal sequence to sequence generation task. In natural language processing and computer vision, tasks like text to speech (TTS) [69, 41, 43, 83] and speech to gesture [22, 28, 23], image/video captioning (pixels to text) [13, 44, 58, 48] involve solving the cross-modal sequence to sequence generation problem. Initially, combination of CNNs and RNNs [86, 85, 91, 93] were prominent in approaching this problem. More recently, with the development of attention mechanism [84], transformer based networks achieve top performance for visual-text [95, 77, 19, 54, 38, 76, 76], visual-audio [26, 89] cross-modal sequence to sequence generation task. Our work explores audio to 3D motion in a transformer based architecture. While all cross-modal problems induce its own challenges, the problem of music to 3D dance is uniquely challenging in that there are many ways to dance to the same music and that the same dance choreography may be used for multiple music. We hope the proposed AIST++ dataset advances research in this relatively under-explored problem.

3. AIST++ Dataset

Data Collection We generate the proposed 3D motion dataset from an existing database called AIST Dance Database [82]. AIST is only a collection of videos without any 3D information. Although it contains multi-view videos of dancers, these cameras are not calibrated, making 3D reconstruction of dancers a non-trivial effort. We recover the camera calibration parameters and the 3D human motion in terms of SMPL parameters. Please find the details of this algorithm in the Appendix. Although we adopt the best practices in reconstructing this data, no code base exist for this particular problem setup and running this pipeline on a large-scale video dataset requires non-trivial amount of compute and effort. We will make the 3D data and camera parameters publicly available, which allows the community to benchmark on this dataset on an equal footing.

Dataset Description Resulting AIST++ is a large-scale 3D human dance motion dataset that contains a wide variety of 3D motion paired with music. It has the following extra annotations for each frame:

- 9 views of camera intrinsic and extrinsic parameters;

Dataset	Music	3D Joint$_{pos}$	3D Joint$_{rot}$	2D Kpt	Views	Images	Genres	Subjects	Sequences	Seconds
AMASS[59]	✗	✓	✓	✗	0	0	0	344	11265	145251
Human3.6M[39]	✗	✓	✓	✓	4	3.6M	0	11	210	71561
Dance with Melody[79]	✓	✓	✗	✗	0	0	4	-	61	5640
GrooveNet [5]	✓	✓	✗	✗	0	0	1	1	2	1380
DanceNet [96]	✓	✓	✗	✗	0	0	2	2	2	3472
EA-MUD [78]	✓	✓	✗	✗	0	0	4	-	17	1254
AIST++	✓	✓	✓	✓	9	10.1M	10	30	1408	18694

Table 1: **3D Dance Datasets Comparisons.** The proposed AIST++ dataset is the largest dataset with 3D dance motion paired with music. We also have the largest variety of subjects and genres. Furthermore, our dataset is the only one that comes with image frames, as other dance datasets only contain motion capture dataset. We include popular 3D motion dataset without any music in the first two rows for reference.

- 17 COCO-format[70] human joint locations in both 2D and 3D;
- 24 SMPL [57] pose parameters along with the global scaling and translation.

Besides the above properties, AIST++ dataset also contains multi-view synchronized image data unlike prior 3D dance dataset, making it useful for other research directions such as 2D/3D pose estimation. To our knowledge, AIST++ is the largest 3D human dance dataset with **1408** sequences, **30** subjects and **10** dance genres with basic and advanced choreographies. See Table. 1 for comparison with other 3D motion and dance datasets. AIST++ is a complementary dataset to existing 3D motion dataset such as AMASS [59], which contains only 17.8 minutes of dance motions with no accompanying music.

Owing to the richness of AIST, AIST++ contains 10 dance genres: Old School (Break, Pop, Lock and Waack) and New School (Middle Hip-hop, LA-style Hip-hop, House, Krump, Street Jazz and Ballet Jazz). Please see the Appendix for more details and statistics. The motions are equally distributed among all dance genres, covering wide variety of music tempos denoted as beat per minute (BPM)[61]. Each genre of dance motions contains 85% of basic choreographies and 15% of advanced choreographies, in which the former ones are those basic short dancing movements while the latter ones are longer movements freely designed by the dancers. However, note that AIST is an instructional database and records multiple dancers dancing the same choreography for different music with varying BPM, a common practice in dance. This posits a unique challenge in cross-modal sequence-to-sequence generation. We carefully construct non-overlapping train and val subsets on AIST++ to make sure neither choreography nor music is shared across the subsets.

4. Music Conditioned 3D Dance Generation

Here we describe our approach towards the problem of music conditioned 3D dance generation. Specifically, given a 2-second seed sample of motion represented as $\mathbf{X} = (x_1, \ldots, x_T)$ and a longer conditioning music sequence represented as $\mathbf{Y} = (y_1, \ldots, y_{T'})$, the problem is to generate a sequence of future motion $\mathbf{X}' = (x_{T+1}, \ldots, x_{T'})$ from time step $T+1$ to T', where $T' \gg T$.

Preliminaries Transformer [84] is an attention based network widely applied in natural language processing. A basic transformer building block (shown in of Figure 3 (a)) has multiple layers with each layer composed of a multi-head attention-layer (Attn) followed by a feed forward layer (FF). The multi-head attention-layer embeds input sequence $\mathbf{X}$ into an internal representation often referred to as the context vector $\mathbf{C}$. Specifically, the output of the attention layer, the context vector $\mathbf{C}$ is computed using the query vector $\mathbf{Q}$ and the key $\mathbf{K}$ value $\mathbf{V}$ pair from input with or without a mask $\mathbf{M}$ via,

$$\begin{aligned}\mathbf{C} &= \text{FF}(\text{Attn}(\mathbf{Q}, \mathbf{K}, \mathbf{V}, \mathbf{M})) \\ &= \text{FF}(\text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T + \mathbf{M}}{\sqrt{D}}\right)\mathbf{V}), \\ \mathbf{Q} &= \mathbf{X}\mathbf{W}^Q, \mathbf{K} = \mathbf{X}\mathbf{W}^K, \mathbf{V} = \mathbf{X}\mathbf{W}^V \end{aligned} \quad (1)$$

where D is the number of channels in the attention layer and $\mathbf{W}$ are trainable weights. The design of the mask function is a key parameter in a transformer. In natural language generation, causal models such as GPT [66] uses an upper triangular look-ahead mask $\mathbf{M}$ to enable causal attention where each token can only look at past inputs. This allows efficient inference at test time, since intermediate context vectors do not need to be recomputed, especially given the large context window in these models (2048). On the other hand, models like BERT [17] employ full-attention for feature learning, but rarely are these models employed in an auto-regressive manner, due to its inefficiency at test time.

4.1. Full Attention Cross-Modal Transformer

We propose Full Attention Cross-Modal Transformer (FACT) model for the task of 3D dance motion generation. Given the seed motion $\mathbf{X}$ and audio features $\mathbf{Y}$, FACT first encodes these inputs using a motion transformer f_{mot} and audio transformer f_{audio} into motion and audio embeddings $\mathbf{h}^x{}_{1:T}$ and $\mathbf{h}^y{}_{1:T'}$ respectively. These are then concatenated

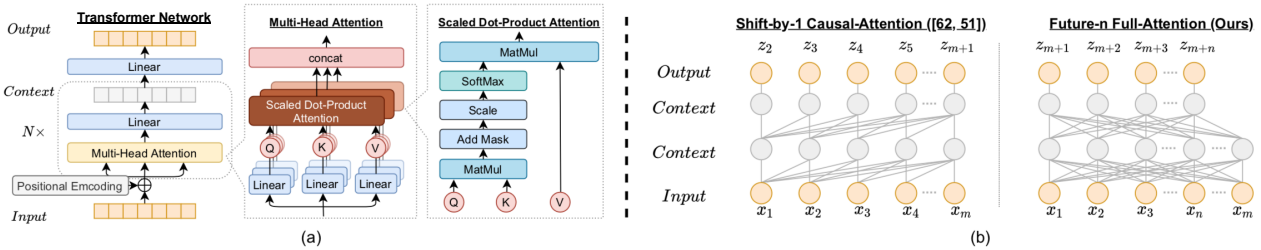

Figure 3: **FACT Model Details.** (a) The structure of the audio/motion/cross-modal transformer with N attention layers. (b) Attention and supervision mechanism as a simplified two-layer model. Models like GPT [66] and the motion generator of [55] use causal attention (left) to predict the immediate next output for each input nodes. We employ full-attention and predict n future from the last input timestamp m (right). The dots on the bottom row are the input tensors, which are computed into context tensors through causal (left) and full (right) attention transformer layer. The output (predictions) are shown on the top. We empirically show that these design choices are critical in generating non-freezing, more realistic motion sequences.

and sent to a cross-modal transformer f_{cross}, which learns the correspondence between both modalities and generates N future motion sequences $\mathbf{X}'$, which is used to train the model in a self-supervised manner. All three transformers are jointly learned in an end-to-end manner. This process is illustrated in Figure 2. At test time, we apply this model in an auto-regressive framework, where we take the first predicted motion as the input of the next generation step and shift all conditioning by one.

FACT involves three key design choices that are critical for producing realistic 3D dance motion from music. First, all of the transformers use full-attention mask. We can still apply this model efficiently in an auto-regressive framework at test time, since our context window is not prohibitively large (240). The full-attention model is more expressive than the causal model because internal tokens have access to all inputs. Due to this full-attention design, we train our model to only predict the unseen future after the context window. In particular, we train our model to predict N futures beyond the current input instead of just 1 future motion. This encourages the network to pay more attention to the temporal context, and we experimentally validate that this is a key factor training a model that does not suffer from motion freezing or diverging after a few generation steps. This attention design is in contrast to prior work that employ transformers for the task of 3D motion [3] or dance generation [55], which applies GPT [66] style causal transformer trained to predict the immediate next future token. We illustrate this difference in Figure 3 (b).

Lastly, we fuse the two embeddings early and employ a deep 12-layer cross-modal transformer module. This is in contrast to prior work that used a single MLP to combine the audio and motion embeddings [55], and we find that deep cross-modal module is essential for training a model that actually pays attention to the input music. This is particularly important as in dance, similar choreography can be used for multiple music. This also happens in AIST dataset, and we find that without a deep cross-modal module, the network is prone to ignoring the conditioning music. We experimentally validate this in Section 5.2.3.

5. Experiments

5.1. AIST++ Motion Quality Validation

We first carefully validate the quality of our 3D motion reconstruction. Possible error sources that may affect the quality of our 3D reconstruction include inaccurate 2D keypoints detection and the estimated camera parameters. As there is no 3D ground-truth for AIST dataset, our validation here is based-on the observation that the re-projected 2D keypoints should be consistent with the predicted 2D keypoints which have high prediction confidence in each image. We use the 2D mean per joint position error MPJPE-2D, commonly used for 3D reconstruction quality measurement [46, 39, 65]) to evaluate the consistency between the predicted 2D keypoints and the reconstructed 3D keypoints along with the estimated camera parameters. Note we only consider 2D keypoints with prediction confidence over 0.5 to avoid noise. The MPJPE-2D of our entire dataset is 6.2 pixels on the 1920×1080 image resolution, and over 86% of those has less than 10 pixels of error. Besides, we also calculate the PCKh metric introduced in [6] on our AIST++. The PCKh@0.5 on the whole set is 98.7%, meaning the reconstructed 3D keypoints are highly consistent with the predicted 2D keypoints. Please refer to the Appendix for detailed analysis of MPJPE-2D and PCKh on AIST++.

5.2. Music Conditioned 3D Motion Generation

5.2.1 Experimental Setup

Dataset Split All the experiments in this paper are conducted on our AIST++ dataset, which to our knowledge is the largest dataset of this kind. We split AIST++ into *train* and *test* set, and report the performance on the *test* set only. We carefully split the dataset to make sure that the music and dance motion in the *test* set does not overlap with that in the *train* set. To build the *test* set, we first select one music piece from each of the 10 genres. Then for each music

	Motion Quality		Motion Diversity		Motion-Music Corr	User Study
	$\text{FID}_k \downarrow$	$\text{FID}_g \downarrow$	$\text{Dist}_k \uparrow$	$\text{Dist}_g \uparrow$	BeatAlign ↑	FACT WinRate↓
AIST++	–	–	9.057	7.556	0.292	–
AIST++ (random)	–	–	–	–	0.213	25.4%
Li *et al*. [55]	86.43	20.58	6.85*	4.93	0.232	80.6%
Dancenet [96]	69.18	17.76	2.86	2.72	0.232	71.1%
DanceRevolution [36]	73.42	31.01	3.52	2.46	0.220	77.0%
FACT (ours)	**35.35**	**12.40**	**5.94**	**5.30**	**0.241**	–

Table 2: **Conditional Motion Generation Evaluation on AIST++ dataset.** Comparing to the three recent state-of-the-art methods, our model generates motions that are more realistic, better correlated with input music and more diversified when conditioned on different music. *Note Li *et al*. [55]'s generated motions are discontinuous making its average kinetic feature distance (FID_k) abnormally high.

piece, we randomly select two dancers, each with two different choreographies paired with that music, resulting in total 40 unique choreographies in the *test* set. The *train* set is built by excluding all test musics and test choreographies from AIST++, resulting in total 329 unique choreographies in the *train* set. Note that in the test set we *intentionally* pick music pieces with different BPMs so that it covers all kinds of BPMs ranging from 80 to 135 in AIST++.

Implementation Details In our main experiment, the input of the model contains a seed motion sequence with 120 frames (2 seconds) and a music sequence with 240 frames (4 seconds), where the two sequences are aligned on the first frame. The output of the model is the future motion sequence with $N = 20$ frames supervised by $L2$ loss. During inference we continually generate future motions in a auto-regressive manner at 60 FPS, where only the first predicted motion is kept in every step. We use the publicly available audio processing toolbox Librosa [60] to extract the music features including: 1-dim *envelope*, 20-dim *MFCC*, 12-dim *chroma*, 1-dim *one-hot peaks* and 1-dim *one-hot beats*, resulting in a 35-dim music feature. We combine the 9-dim rotation matrix representation for all 24 joints, along with a 3-dim global translation vector, resulting in a 219-dim motion feature. Both these raw audio and motion features are first embedded into 800-dim hidden representations with linear layers, then added with learnable positional encoding, before they were input into the transformer layers. All the three (audio, motion, cross-modal) transformers have 10 attention heads with 800 hidden size. The number of attention layers in each transformer varies based on the experiments, as described in Sec. 5.2.3. We disregard the last linear layer in the audio/motion transformer and the positional encoding in the cross-modal transformer, as they are not necessary in the FACT model. All our experiments are trained with 16 batch size using Adam [45] optimizer. The learning rate starts from $1\text{e}{-4}$ and drops to $\{1\text{e}{-5}, 1\text{e}{-6}\}$ after $\{60k, 100k\}$ steps. The training finishes after $300k$, which takes 3 days on 4 TPUs. For baselines, we compare with the latest work on 3D dance generation that take music and seed motion as input, including Dancenet [96] and Li *et al*. [55]. For a more comprehensive evaluation we also compare with the recent state-of-the-art 2D dance generation method DanceRevolution [36]. We adapt this work to output 3D joint locations which can be directly compared with our results quantitatively, though joint locations do not allow immediate re-targeting. We train and test these baselines on the same dataset with ours using the *official* code provided by the authors.

5.2.2 Quantitative Evaluation

In this section, we evaluate our proposed model FACT on the following aspects: (1) motion quality, (2) generation diversity and (3) motion-music correlation. Experiments results (shown in Table 2) show that our model out-performs state-of-the-art methods [55, 36, 96], on those criteria.

Motion Quality Similar to prior works [55, 36], we evaluate the generated motion quality by calculating the distribution distance between the generated and the ground-truth motions using Frechet Inception Distance (FID) [32] on the extracted motion features. As prior work used motion-encoders that are not public, we measure FID with two well-designed motion feature extractors [62, 63] implemented in *fairmotion* [29]: (1) a geometric feature extractor that produces a boolean vector $\mathbf{z}_g \in \mathbb{R}^{33}$ expressing geometric relations between certain body points in the motion sequence $X \in \mathbb{R}^{T\times N\times 3}$, (2) a kinetic feature extractor [63] that maps a motion sequence X to $\mathbf{z}_k \in \mathbb{R}^{72}$, which represents the kinetic aspects of the motion such as velocity and accelerations. We denote the FID based on these geometric and kinetic features as $\text{FID}_g$and FID_k, respectively. The metrics are calculated between the real dance motion sequences in AIST++ test set and 40 generated motion sequences each with $T = 1200$ frames (20 secs). As shown in Table 2, our generated motion sequences have a much closer distribution to ground-truth motions compared with the three baselines. We also visualize the generated sequences from the baselines in our supplemental video.

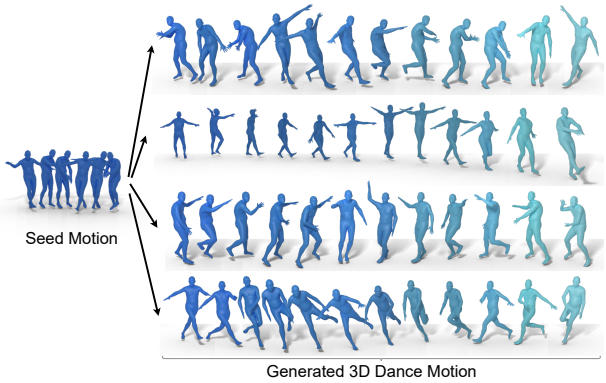

Figure 4: **Diverse Generation Results.** Here we visualize 4 different dance motions generated using *different* music but the *same* seed motion. On the left we illustrate the 2 second seed motion and on the right we show the generated 3D dance sequences subsampled by 2 seconds. For rows top to bottom, the genres of the conditioning music are: Break, Ballet Jazz, Krump and Middle Hip-hop. Note that the seed motion come from hip-hop dance.Our model is able to adapt the dance style when given a more modern dance music (second row: Ballet Jazz). Please see more results in the supplementary video.

Generation Diversity We also evaluate our model's ability to generate diverse dance motions when given various input music compared with the baseline methods. Similar to the prior work [36], we calculate the average Euclidean distance in the feature space across 40 generated motions on the AIST++ *test* set to measure the diversity. The motion diversity in the geometric feature space and in the kinetic feature space are noted as $\text{Dist}_m$and Dist_k, respectively. Table 2 shows that our method generates more diverse dance motions comparing to the baselines except Li *et al*. [55], which discretizes the motion, leading to discontinuous outputs that results in high Dist_k. Our generated diverse motions are visualized in Figure 4.

Motion-Music Correlation Further, we evaluate how much the generated 3D motion correlates to the input music. As there is no well-designed metric to measure this property, we propose a novel metric, Beat Alignment Score (BeatAlign), to evaluate the motion-music correlation in terms of the similarity between the kinematic beats and music beats. The music beats are extracted using *librosa* [60] and the kinematic beats are computed as the local minima of the kinetic velocity, as shown in Figure 5. The Beat Alignment Score is then defined as the average distance between every kinematic beat and its nearest music beat. Specifically, our Beat Alignment Score is defined as:

$$\text{BeatAlign} = \frac{1}{m}\sum_{i=1}^{m}\exp\left(-\frac{\min_{\forall t_j^y \in B^y}||t_i^x - t_j^y||^2}{2\sigma^2}\right) \quad (2)$$

where $B^x = \{t_i^x\}$ is the kinematic beats, $B^y = \{t_j^y\}$ is the music beats and σ is a parameter to normalize sequences

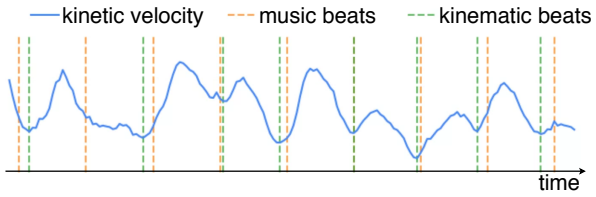

Figure 5: **Beats Alignment between Music and Generated Dance.** Here we visualize the kinetic velocity (blue curve) and kinematic beats (green dotted line) of our generated dance motion, as well as the music beats (orange dotted line). The kinematic beats are extracted by finding local minima from the kinetic velocity curve.

with different FPS. We set $\sigma = 3$ in all our experiments as the FPS of all our experiments sequences is 60. A similar metric Beat Hit Rate was introduced in [51, 36], but this metric requires a dataset dependent handcrafted threshold to decide the alignment ("hit") while ours directly measure the distances. This metric is explicitly designed to be uni-directional as dance motion does not necessarily *have to* match with every music beat. On the other hand, every kinetic beat is expected to have a corresponding music beat. To calibrate the results, we compute the correlation metrics on the entire AIST++ dataset (upper bound) and on the random-paired data (lower bound). As shown in Table 2, our generated motion is better correlated with the input music compared to the baselines. We also show one example in Figure 5 that the kinematic beats of our generated motion align well with the music beats. However, when comparing to the real data, all four methods including ours have a large space for improvement. This reflects that music-motion correlation is still a challenging problem.

5.2.3 Ablation Study

We conduct the following ablation experiments to study the effectiveness of our key design choices: Full-Attention Future-N supervision, and early cross-modal fusion. Please refer to our supplemental video for qualitative comparison. The effectiveness of different model architectures is measured quantitatively using the motion quality (FID_k, FID_g) and the music-motion correlation (BeatAlign) metrics, as shown in Table 4 and Table 3.

Full-Attention Future-N Supervision Here we dive deep into the attention mechanism and our future-N supervision scheme. We set up four different settings: causal-attention shift-by-1 supervision, and full-attention with future-$\{1, 10, 20\}$ supervision. Qualitatively, we find that the motion generated by the causal-attention with shift-by-1 supervision (as done in [55, 66, 3]) starts to freeze after several seconds (please see the supplemental video). Similar problem was reported in the results of [3]. Quantitatively (shown in the Table 3), when using causal-attention shift-by-1 supervision, the FIDs are large meaning that the difference between generated and ground-truth motion se-

Attn-Supervision	$FID_k \downarrow$	$FID_g \downarrow$	BeatAlign ↑
Causal-Attn-Shift-by-1	111.69	21.43	0.217
Full-Attn-F1 (FACT-1)	207.74	19.35	0.233
Full-Attn-F10 (FACT-10)	**35.10**	15.17	0.239
Full-Attn-F20 (FACT-20)	35.35	**12.39**	**0.241**

Table 3: **Ablation Study on Attention and Supervision Mechanism.** Causal-attention shift-by-1 supervision tends to generate freezing motions in the long-term. While Full-attention supervised more future frames boost the ability of generating more realistic dance motions.

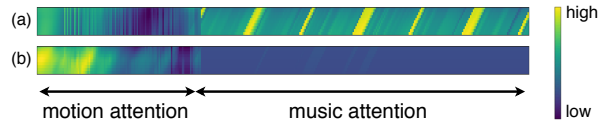

Figure 6: **Attention Weights Visualization.** We compare the attention weights from the last layer of the (a) 12-layer cross-modal transformer and (b) 1-layer cross-modal transformer. Deeper cross-modal transformer pays equal attention to motion and music, while a shallower one pays more attention to motion.

quences is substantial. For the full-attention with future-1 supervision setting, the results rapidly drift during long-range generation. However, when the model is supervised with 10 or 20 future frames, it pays more attention to the temporal context. Thus, it learns to generate good quality (non-freezing, non-drifting) long-range motion.

Early Cross-Modal Fusion Here we investigate when to fuse the two input modalities. We conduct experiments in three settings, (1) *No-Fusion*: 14-layer motion transformer only; (2) *Late-Fusion*: 13-layer motion/audio transformer with 1-layer cross-modal transformer; (3) *Early-Fusion*: 2-layer motion/audio transformer with 12-layer cross-modal transformer. For fair comparison, we change the number of attention layers in the motion/audio transformer and the cross-modal transformer but keep the total number of the attention layers fixed. Table 4 shows that the early fusion between two input modalities is critical to generate motions that are well correlated with input music. Also we show in Figure 6 that Early-Fusion allows the cross-model transformer pays more attention to the music, while Late-Fusion tend to ignore the conditioning music. This also aligns with our intuition that the two modalities need to be fully fused for better cross-modal learning, as contrast to prior work that uses a single MLP to combine the audio and motion [55].

5.2.4 User Study

Finally, we perceptually evaluate the motion-music correlation with a user study to compare our method with the three baseline methods and the "random" baseline, which randomly combines AIST++ motion-music. (Refer to the Appendix for user study details.) In this study, each user

Cross-Modal Fusion	$FID_k \downarrow$	$FID_g \downarrow$	BeatAlign ↑
No-Fusion	45.66	13.27	0.228*
Late-Fusion	45.76	14.30	0.234
Early-Fusion	**35.35**	**12.39**	**0.241**

Table 4: **Ablation Study on Cross-modal Fusion.** Early fusion of the two modalities allows the model to generate motion sequences align better with the conditioning music. *Note this number is calculated using the music paired with the input motion.

is asked to watch 10 videos showing one of our results and one random counterpart, and answer the question *"which person is dancing more to the music? LEFT or RIGHT"* for each video. For user study on each of the four baselines, we invite 30 participants, ranging from professional dancers to people who rarely dance. We analyze the feedback and the results are: (1) 81% of our generated dance motion is better than Li *et al*. [55]; (2) 71% of our generated dance motion is better than Dancenet [96]; (3) 77% of our generated dance motion is better than DanceRevolution [36]; (4) 75% of the unpaired AIST++ dance motion is better than ours. Clearly we surpass the baselines in the user study. But because the "random" baseline consists of real advanced dance motions that are extremely expressive, participants are biased to prefer it over ours. However, quantitative metrics show that our generated dance is more aligned with music.

6. Conclusion and Discussion

In this paper, we present a cross-modal transformer-based neural network architecture that can not only learn the audio-motion correspondence but also can generate non-freezing high quality 3D motion sequences conditioned on music. We also construct the largest 3D human dance dataset: AIST++. This proposed, multi-view, multi-genre, cross-modal 3D motion dataset can not only help research in the conditional 3D motion generation research but also human understanding research in general. While our results shows a promising direction in this problem of music conditioned 3D motion generation, there are more to be explored. First, our approach is kinematic based and we do not reason about physical interactions between the dancer and the floor. Therefore the global translation can lead to artifacts such as foot sliding and floating. Second, our model is currently deterministic. Exploring how to generate multiple realistic dance per music is an exciting direction.

7. Acknowledgement

We thank Chen Sun, Austin Myers, Bryan Seybold and Abhijit Kundu for helpful discussions. We thank Emre Aksan and Jiaman Li for sharing their code. We also thank Kevin Murphy for the early attempts on this direction, as well as Peggy Chi and Pan Chen for the help on user study experiments.

References

[1] Mixamo. https://www.mixamo.com/. 1, 3

[2] Hyemin Ahn, Jaehun Kim, Kihyun Kim, and Songhwai Oh. Generative autoregressive networks for 3d dancing move synthesis from music. *IEEE Robotics and Automation Letters*, 5(2):3500–3507, 2020. 3

[3] Emre Aksan, Peng Cao, Manuel Kaufmann, and Otmar Hilliges. Attention, please: A spatio-temporal transformer for 3d human motion prediction. *arXiv preprint arXiv:2004.08692*, 2020. 2, 3, 5, 7

[4] Emre Aksan, Manuel Kaufmann, and Otmar Hilliges. Structured prediction helps 3d human motion modelling. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 7144–7153, 2019. 2, 3

[5] Omid Alemi, Jules Françoise, and Philippe Pasquier. Groovenet: Real-time music-driven dance movement generation using artificial neural networks. *networks*, 8(17):26, 2017. 3, 4

[6] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *Proceedings of the IEEE Conference on computer Vision and Pattern Recognition*, pages 3686–3693, 2014. 5

[7] Okan Arikan and David A Forsyth. Interactive motion generation from examples. *ACM Transactions on Graphics (TOG)*, 21(3):483–490, 2002. 2

[8] Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. Scheduled sampling for sequence prediction with recurrent neural networks. In *Advances in neural information processing systems*, 2015. 3

[9] Uttaran Bhattacharya, Nicholas Rewkowski, Abhishek Banerjee, Pooja Guhan, Aniket Bera, and Dinesh Manocha. Text2gestures: A transformer-based network for generating emotive body gestures for virtual agents. *arXiv preprint arXiv:2101.11101*, 2021. 3

[10] Richard Bowden. Learning statistical models of human motion. In *IEEE Workshop on Human Modeling, Analysis and Synthesis, CVPR*, volume 2000, 2000. 2

[11] Matthew Brand and Aaron Hertzmann. Style machines. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, pages 183–192, 2000. 2

[12] Judith Bütepage, Michael J Black, Danica Kragic, and Hedvig Kjellström. Deep representation learning for human motion prediction and classification. In *CVPR*, page 2017, 2017. 3

[13] Slides by Saheel. Baby talk: Understanding and generating image descriptions. 3

[14] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields. In *arXiv preprint arXiv:1812.08008*, 2018. 3

[15] Kang Chen, Zhipeng Tan, Jin Lei, Song-Hai Zhang, Yuan-Chen Guo, Weidong Zhang, and Shi-Min Hu. Choreomaster: choreography-oriented music-driven dance synthesis. *ACM Transactions on Graphics (TOG)*, 40(4):1–13, 2021. 3

[16] Hsu-kuang Chiu, Ehsan Adeli, Borui Wang, De-An Huang, and Juan Carlos Niebles. Action-agnostic human pose forecasting. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1423–1432. IEEE, 2019. 3

[17] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 4

[18] Xiaoxiao Du, Ram Vasudevan, and Matthew Johnson-Roberson. Bio-lstm: A biomechanically inspired recurrent neural network for 3-d pedestrian pose and gait prediction. *IEEE Robotics and Automation Letters*, 4(2):1501–1508, 2019. 3

[19] Xuguang Duan, Wenbing Huang, Chuang Gan, Jingdong Wang, Wenwu Zhu, and Junzhou Huang. Weakly supervised dense event captioning in videos. In *Advances in Neural Information Processing Systems*, pages 3059–3069, 2018. 3

[20] Rukun Fan, Songhua Xu, and Weidong Geng. Example-based automatic music-driven conventional dance motion synthesis. *IEEE transactions on visualization and computer graphics*, 18(3):501–515, 2011. 3

[21] Joao P Ferreira, Thiago M Coutinho, Thiago L Gomes, José F Neto, Rafael Azevedo, Renato Martins, and Erickson R Nascimento. Learning to dance: A graph convolutional adversarial network to generate realistic dance motions from audio. *Computers & Graphics*, 94:11–21. 3

[22] Ylva Ferstl and Rachel McDonnell. Investigating the use of recurrent motion modelling for speech gesture generation. In *Proceedings of the 18th International Conference on Intelligent Virtual Agents*, pages 93–98, 2018. 3

[23] Ylva Ferstl, Michael Neff, and Rachel McDonnell. Adversarial gesture generation with realistic gesture phasing. *Computers & Graphics*, 2020. 3

[24] Katerina Fragkiadaki, Sergey Levine, Panna Felsen, and Jitendra Malik. Recurrent network models for human dynamics. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4346–4354, 2015. 3

[25] Aphrodite Galata, Neil Johnson, and David Hogg. Learning variable-length markov models of behavior. *Computer Vision and Image Understanding*, 81(3):398–413, 2001. 2

[26] Chuang Gan, Deng Huang, Peihao Chen, Joshua B Tenenbaum, and Antonio Torralba. Foley music: Learning to generate music from videos. *arXiv preprint arXiv:2007.10984*, 2020. 3

[27] Partha Ghosh, Jie Song, Emre Aksan, and Otmar Hilliges. Learning human motion models for long-term predictions. In *2017 International Conference on 3D Vision (3DV)*, pages 458–466. IEEE, 2017. 3

[28] Shiry Ginosar, Amir Bar, Gefen Kohavi, Caroline Chan, Andrew Owens, and Jitendra Malik. Learning individual styles of conversational gesture. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3497–3506, 2019. 3

[29] Deepak Gopinath and Jungdam Won. fairmotion - tools to load, process and visualize motion capture data. Github, 2020. 6

[30] Rachel Heck and Michael Gleicher. Parametric motion graphs. In *Proceedings of the 2007 symposium on Interactive 3D graphics and games*, pages 129–136, 2007. 2

[31] Alejandro Hernandez, Jurgen Gall, and Francesc Moreno-Noguer. Human motion prediction via spatio-temporal inpainting. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 7134–7143, 2019. 2

[32] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in neural information processing systems*, pages 6626–6637, 2017. 6

[33] Daniel Holden, Taku Komura, and Jun Saito. Phase-functioned neural networks for character control. *ACM Transactions on Graphics (TOG)*, 36(4):1–13, 2017. 3

[34] Daniel Holden, Jun Saito, and Taku Komura. A deep learning framework for character motion synthesis and editing. *ACM Transactions on Graphics (TOG)*, 35(4):1–11, 2016. 2

[35] Daniel Holden, Jun Saito, Taku Komura, and Thomas Joyce. Learning motion manifolds with convolutional autoencoders. In *SIGGRAPH Asia 2015 Technical Briefs*, pages 1–4. 2015. 2

[36] Ruozi Huang, Huang Hu, Wei Wu, Kei Sawada, Mi Zhang, and Daxin Jiang. Dance revolution: Long-term dance generation with music via curriculum learning. In *International Conference on Learning Representations*, 2021. 3, 6, 7, 8

[37] Yinghao Huang, Manuel Kaufmann, Emre Aksan, Michael J. Black, Otmar Hilliges, and Gerard Pons-Moll. Deep inertial poser learning to reconstruct human pose from sparseinertial measurements in real time. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 37(6):185:1–185:15, Nov. 2018. 3

[38] Vladimir Iashin and Esa Rahtu. Multi-modal dense video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 958–959, 2020. 3

[39] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7):1325–1339, jul 2014. 3, 4, 5

[40] Ashesh Jain, Amir R Zamir, Silvio Savarese, and Ashutosh Saxena. Structural-rnn: Deep learning on spatio-temporal graphs. In *Proceedings of the ieee conference on computer vision and pattern recognition*, pages 5308–5317, 2016. 3

[41] Ye Jia, Melvin Johnson, Wolfgang Macherey, Ron J Weiss, Yuan Cao, Chung-Cheng Chiu, Naveen Ari, Stella Laurenzo, and Yonghui Wu. Leveraging weakly supervised data to improve end-to-end speech-to-text translation. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7180–7184. IEEE, 2019. 3

[42] Hsuan-Kai Kao and Li Su. Temporally guided music-to-body-movement generation. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 147–155, 2020. 3

[43] Shigeki Karita, Nanxin Chen, Tomoki Hayashi, Takaaki Hori, Hirofumi Inaguma, Ziyan Jiang, Masao Someki, Nelson Enrique Yalta Soplin, Ryuichi Yamamoto, Xiaofei Wang, et al. A comparative study on transformer vs rnn in speech applications. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 449–456. IEEE, 2019. 3

[44] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137, 2015. 3

[45] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[46] Muhammed Kocabas, Nikos Athanasiou, and Michael J. Black. Vibe: Video inference for human body pose and shape estimation, 2019. 5

[47] Lucas Kovar, Michael Gleicher, and Frédéric Pighin. Motion graphs. In *ACM SIGGRAPH 2008 classes*, pages 1–10. 2008. 2

[48] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *Proceedings of the IEEE international conference on computer vision*, pages 706–715, 2017. 3

[49] Jogendra Nath Kundu, Himanshu Buckchash, Priyanka Mandikal, Anirudh Jamkhandi, Venkatesh Babu RADHAKRISHNAN, et al. Cross-conditioned recurrent networks for long-term synthesis of inter-person human motion interactions. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2724–2733, 2020. 3

[50] Kimerer LaMothe. The dancing species: how moving together in time helps make us human. *Aeon*, June 2019. 1

[51] Hsin-Ying Lee, Xiaodong Yang, Ming-Yu Liu, Ting-Chun Wang, Yu-Ding Lu, Ming-Hsuan Yang, and Jan Kautz. Dancing to music, 2019. 3, 7

[52] Juheon Lee, Seohyun Kim, and Kyogu Lee. Listen to dance: Music-driven choreography generation using autoregressive encoder-decoder network. *arXiv preprint arXiv:1811.00818*, 2018. 3

[53] Jehee Lee and Sung Yong Shin. A hierarchical approach to interactive motion editing for human-like figures. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, pages 39–48, 1999. 2

[54] Guang Li, Linchao Zhu, Ping Liu, and Yi Yang. Entangled transformer for image captioning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 8928–8937, 2019. 3

[55] Jiaman Li, Yihang Yin, Hang Chu, Yi Zhou, Tingwu Wang, Sanja Fidler, and Hao Li. Learning to generate diverse dance motions with transformer. *arXiv preprint arXiv:2008.08171*, 2020. 3, 5, 6, 7, 8

[56] Zimo Li, Yi Zhou, Shuangjiu Xiao, Chong He, Zeng Huang, and Hao Li. Auto-conditioned recurrent networks for extended complex human motion synthesis. *ICLR*, 2018. 3

[57] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *SIGGRAPH Asia*, 2015. 4

[58] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Neural baby talk. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7219–7228, 2018. 3

[59] Naureen Mahmood, Nima Ghorbani, Nikolaus F Troje, Gerard Pons-Moll, and Michael J Black. Amass: Archive of motion capture as surface shapes. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5442–5451, 2019. 3, 4

[60] 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 *Proceedings of the 14th python in science conference*, volume 8, 2015. 6, 7

[61] Dirk Moelants. Dance music, movement and tempo preferences. In *Proceedings of the 5th Triennial ESCOM Conference*, pages 649–652. Hanover University of Music and Drama, 2003. 4

[62] Meinard Müller, Tido Röder, and Michael Clausen. Efficient content-based retrieval of motion capture data. In *ACM SIGGRAPH 2005 Papers*, pages 677–685. 2005. 6

[63] Kensuke Onuma, Christos Faloutsos, and Jessica K Hodgins. Fmdistance: A fast and effective distance function for motion capture data. In *Eurographics (Short Papers)*, pages 83–86, 2008. 6

[64] Katherine Pullen and Christoph Bregler. Animating by multi-level sampling. In *Proceedings Computer Animation 2000*, pages 36–42. IEEE, 2000. 2

[65] Haibo Qiu, Chunyu Wang, Jingdong Wang, Naiyan Wang, and Wenjun Zeng. Cross view fusion for 3d human pose estimation. In *International Conference on Computer Vision (ICCV)*, 2019. 5

[66] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training, 2018. 4, 5, 7

[67] Xuanchi Ren, Haoran Li, Zijian Huang, and Qifeng Chen. Music-oriented dance video synthesis with pose perceptual loss. *arXiv preprint arXiv:1912.06606*, 2019. 3

[68] Xuanchi Ren, Haoran Li, Zijian Huang, and Qifeng Chen. Self-supervised dance video synthesis conditioned on music. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 46–54, 2020. 3

[69] Yi Ren, Yangjun Ruan, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. Fastspeech: Fast, robust and controllable text to speech. In *Advances in Neural Information Processing Systems*, pages 3171–3180, 2019. 3

[70] Matteo Ruggero Ronchi and Pietro Perona. Benchmarking and error diagnosis in multi-instance pose estimation. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017. 4

[71] Takaaki Shiratori, Atsushi Nakazawa, and Katsushi Ikeuchi. Dancing-to-music character animation. In *Computer Graphics Forum*, volume 25, pages 449–458. Wiley Online Library, 2006. 3

[72] Eli Shlizerman, Lucio Dery, Hayden Schoen, and Ira Kemelmacher-Shlizerman. Audio to body dynamics. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7574–7583, 2018. 3

[73] Sebastian Starke, He Zhang, Taku Komura, and Jun Saito. Neural state machine for character-scene interactions. *ACM Trans. Graph.*, 38(6):209–1, 2019. 3

[74] Sebastian Starke, Yiwei Zhao, Taku Komura, and Kazi Zaman. Local motion phases for learning multi-contact character movements. *ACM Transactions on Graphics (TOG)*, 39(4):54–1, 2020. 3

[75] Statista. https://www.statista.com/statistics/249396/top-youtube-videos-views/, 2020. Accessed: 2020-11-09. 1

[76] Chen Sun, Fabien Baradel, Kevin Murphy, and Cordelia Schmid. Learning video representations using contrastive bidirectional transformer. *arXiv preprint arXiv:1906.05743*, 2019. 3

[77] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. Videobert: A joint model for video and language representation learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 7464–7473, 2019. 3

[78] 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. 3, 4

[79] Taoran Tang, Jia Jia, and Hanyang Mao. Dance with melody: An lstm-autoencoder approach to music-oriented dance synthesis. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 1598–1606, 2018. 3, 4

[80] Graham W Taylor and Geoffrey E Hinton. Factored conditional restricted boltzmann machines for modeling motion style. In *Proceedings of the 26th annual international conference on machine learning*, pages 1025–1032, 2009. 2

[81] Purva Tendulkar, Abhishek Das, Aniruddha Kembhavi, and Devi Parikh. Feel the music: Automatically generating a dance for an input song. *arXiv preprint arXiv:2006.11905*, 2020. 3

[82] Shuhei Tsuchida, Satoru Fukayama, Masahiro Hamasaki, and Masataka Goto. Aist dance video database: Multi-genre, multi-dancer, and multi-camera database for dance information processing. In *Proceedings of the 20th International Society for Music Information Retrieval Conference, ISMIR 2019*, pages 501–510, Delft, Netherlands, Nov. 2019. 2, 3

[83] Jean-Marc Valin and Jan Skoglund. Lpcnet: Improving neural speech synthesis through linear prediction. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5891–5895. IEEE, 2019. 3

[84] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. 3, 4

[85] Subhashini Venugopalan, Marcus Rohrbach, Jeffrey Donahue, Raymond Mooney, Trevor Darrell, and Kate Saenko. Sequence to sequence-video to text. In *Proceedings of the IEEE international conference on computer vision*, pages 4534–4542, 2015. 3

[86] Subhashini Venugopalan, Huijuan Xu, Jeff Donahue, Marcus Rohrbach, Raymond Mooney, and Kate Saenko. Translating videos to natural language using deep recurrent neural networks. *arXiv preprint arXiv:1412.4729*, 2014. 3

[87] Ruben Villegas, Jimei Yang, Duygu Ceylan, and Honglak Lee. Neural kinematic networks for unsupervised motion retargetting. In *CVPR*, 2018. 3

[88] Borui Wang, Ehsan Adeli, Hsu-kuang Chiu, De-An Huang, and Juan Carlos Niebles. Imitation learning for human pose prediction. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 7124–7133, 2019. 3

[89] Haoming Xu, Runhao Zeng, Qingyao Wu, Mingkui Tan, and Chuang Gan. Cross-modal relation-aware networks for audio-visual event localization. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 3893–3901, 2020. 3

[90] Nelson Yalta, Shinji Watanabe, Kazuhiro Nakadai, and Tetsuya Ogata. Weakly-supervised deep recurrent neural networks for basic dance step generation. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2019. 3

[91] Li Yao, Atousa Torabi, Kyunghyun Cho, Nicolas Ballas, Christopher Pal, Hugo Larochelle, and Aaron Courville. Describing videos by exploiting temporal structure. In *Proceedings of the IEEE international conference on computer vision*, pages 4507–4515, 2015. 3

[92] Zijie Ye, Haozhe Wu, Jia Jia, Yaohua Bu, Wei Chen, Fanbo Meng, and Yanfeng Wang. Choreonet: Towards music to dance synthesis with choreographic action unit. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 744–752, 2020. 3

[93] Haonan Yu, Jiang Wang, Zhiheng Huang, Yi Yang, and Wei Xu. Video paragraph captioning using hierarchical recurrent neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4584–4593, 2016. 3

[94] He Zhang, Sebastian Starke, Taku Komura, and Jun Saito. Mode-adaptive neural networks for quadruped motion control. *ACM Transactions on Graphics (TOG)*, 37(4):1–11, 2018. 3

[95] Luowei Zhou, Yingbo Zhou, Jason J Corso, Richard Socher, and Caiming Xiong. End-to-end dense video captioning with masked transformer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8739–8748, 2018. 3

[96] Wenlin Zhuang, Congyi Wang, Siyu Xia, Jinxiang Chai, and Yangang Wang. Music2dance: Music-driven dance generation using wavenet. *arXiv preprint arXiv:2002.03761*, 2020. 3, 4, 6, 8

[97] Wenlin Zhuang, Yangang Wang, Joseph Robinson, Congyi Wang, Ming Shao, Yun Fu, and Siyu Xia. Towards 3d dance motion synthesis and control. *arXiv preprint arXiv:2006.05743*, 2020. 3