The Thirty-Eighth AAAI Conference on Artificial Intelligence (AAAI-24)

Everything2Motion: Synchronizing Diverse Inputs via a Unified Framework for Human Motion Synthesis

Zhaoxin Fan[1]*, Longbin Ji[2]*, Pengxin Xu[1], Fan Shen[1], Kai Chen[3]
[1]Psyche AI Inc
[2]Xi'an Jiaotong Liverpool University
[3]The Hong Kong University of Science and Technology
fanzhaoxin@ruc.edu.cn, robingg1100@gmail.com

Abstract

In the dynamic field of film and game development, the emergence of human motion synthesis methods has revolutionized avatar animation. Traditional methodologies, typically reliant on single modality inputs like text or audio, employ modality-specific model frameworks, posing challenges for unified model deployment and application. To address this, we propose Everything2Motion, a unified model framework. Everything2Motion consists of three key modules. The Input-Output Modality Modulation module tailors structures for specific multimodal inputs, eliminating the need for modality-specific frameworks. The Query-aware Autoencoder, based on the transformer encoder-decoder architecture, enables efficient latent motion generation. Lastly, the Prior Motion Distillation Decoder, a pretrained module, enhances the final skeleton sequence's naturalness and fluidity. Comprehensive experiments on several public datasets demonstrate the effectiveness of Everything2Motion, highlighting its potential for practical applications and setting a new benchmark in human motion synthesis.

Introduction

The task of synthesizing realistic human motions — a cornerstone in the animation of robots (Khatib et al. 2009) and digital avatars (Lee et al. 2002) — is an imperative at the crossroads of various disciplines such as film making (Huang et al. 2019), video game development (Menache 2000), and live broadcasting (Fan et al. 2022). The pursuit of authenticity in digital human movements, particularly those that exhibit a nuanced response to environmental stimuli, has traditionally been under the purview of action directors and artists, facilitated by sophisticated equipment. Despite remarkable strides made using these methods, the associated expenses are often prohibitive, particularly for engines tasked with continuous, long-term motion synthesis. This economic challenge has catalyzed the need for alternative approaches that efficiently generate human-like motions given suitable input guidance. Not only would such methods be more economical, but they would also streamline the process of designing character movements across an array of practical applications.

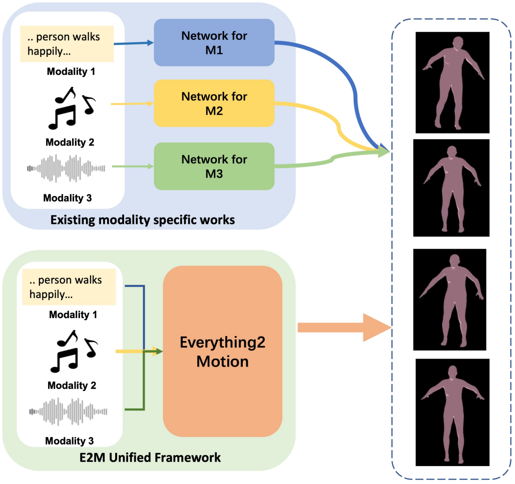

Figure 1: The framework of Everything2Motion. In contrast to previous modality specific works, we design a unified framework for diverse input modalities.

Fortunately, the rapid advancements in deep learning have made it possible to synthesize human motion in a cost-effective and real-time manner (Holden, Saito, and Komura 2016). Two tasks that have particularly benefited from these advancements are: text-to-motion (Guo et al. 2022a,b; Petrovich, Black, and Varol 2022) and music-to-dance (Sun et al. 2020; Zhuang et al. 2022; Li et al. 2021) synthesis. In the former, a sequence of human motions is generated based on textual descriptions, while in the latter, a dance motion sequence is produced corresponding to a given piece of music. Benefiting from variational autoencoder (Petrovich, Black, and Varol 2021; Yan et al. 2018), generative adversarial networks (Barsoum, Kender, and Liu 2018; Liu et al. 2019), diffusion models (Dabral et al. 2023; Zhang et al. 2022), et al., these tasks have made significant strides in terms of progress and performance. The ability to transform textual or musical input into realistic human motion sequences not only demonstrates the power of deep learning but also opens up new avenues for efficient and dynamic character animation.

* Equal Contribution

Despite these advancements, a pertinent issue is that most existing algorithms for text-to-motion and music-to-dance synthesis operate in isolation; different tasks are addressed by designing task-specific models, with no unified framework encompassing motion sequence synthesis under different modalities, even though the output across these tasks is highly similar. This lack of uniformity in model architectures poses a significant challenge to their practical application. For instance, animation production platforms often deploy these tasks within a single application. In such platforms, it is desirable to input data of different modalities, such as music, text, or video, and output corresponding motion sequences. Implementing a different deep learning model for each input modality introduces significant challenges in terms of environment configuration, resource allocation, and underlying software design. Consequently, we pose the question: Is it possible to develop a single network and training framework that can accommodate multiple input modalities?

In response to the aforementioned challenge, we present Everything2Motion, a novel, lightweight framework in this work, as shown in Fig. 1. This unified structure is designed to handle diverse input modalities, creating a consistent protocol for human motion synthesis. Inspired by the multi-modal perceiving capabilities of Perceiver-IO (Jaegle et al. 2021), Everything2Motion leverages a Transformer-based autoencoder architecture specifically designed for uniform motion generation. Firstly, Everything2Motion employs an *Input-Output Modality Modulation (IOM) module*. This module utilizes flexible Residual Convolutional blocks and Self-attention mechanisms that cater to inputs with varying attributes. The IOM module plays a crucial role in transforming the input modality into a shared latent space, serving as a bridge that connects diverse input modalities and ensures that they can be processed using a unified approach. Secondly, a *Query-aware Autoencoder (QA) module* is incorporated. This module, based on the transformer encoder-decoder architecture, initializes a latent query and an output query. It then applies Cross-attention between the latent and output query to extract motion-related information from the latent features. The Query-aware Autoencoder is pivotal in efficiently generating the latent motion representation, thereby providing a critical step in the transformation from input to motion. Lastly, a *Prior Motion Distillation Decoder (PDA-VR) module* is deployed. This module, pre-trained on an extensive motion dataset, refines the output from the Query-aware Autoencoder and maps it to a real-length motion sequence. The Prior Motion Distillation Decoder enhances the smoothness and authenticity of the generated motion, enabling the creation of more realistic and natural-looking movements. Everything2Motion is a specialized framework for unifying diverse input modalities. It only requires simple adaptations to the input, eliminating the need to design a separate deep learning model for each input modality. This unified approach alleviates common challenges encountered during model deployment, such as environment configuration, resource allocation, and underlying software design.

Our model is rigorously evaluated on both text-to-motion and music-to-dance tasks, and benchmarked against previous state-of-the-art methods. The comprehensive experiments conducted on several public datasets underscore the effectiveness of Everything2Motion, demonstrating its potential for real-world applications and setting a new standard in human motion synthesis.

In general, our contributions are summarized: 1)We propose Everything2Motion, a novel framework adept at unifying diverse input modalities for human motion synthesis, thereby eliminating the need for distinct deep learning models for each modality. 2)In our Everything2Motion framework, we introduce a novel triad of components: an Input-Output Modality Modulation module, a Query-aware Autoencoder, and a Prior Motion Distillation Decoder. This innovative architecture demonstrates superior performance in human motion synthesis tasks. 3) To substantiate the effectiveness of our proposed Everything2Motion framework, we undertake comprehensive empirical evaluations across multiple public datasets, focusing on two tasks: music-to-dance and text-to-motion synthesis.

Related Work

Text to Motion Methods

The field of human related machine learning has witnessed rapid advancements, particularly in synthesizing motion from textual descriptions. Pioneering methodologies have emerged, each contributing unique strategies to this domain. JL2P (Ahuja and Morency 2019) leveraged an end-to-end curriculum learning approach and a GRU-based motion decoder to create a joint embedding space of language and pose. Further enhancements were made by (Zhou and Wang 2023) through a hierarchical two-stream sequential model, improving the correspondence between text descriptions and pose sequences. The focus shifted to diversity and detail with TEMOS (Petrovich, Black, and Varol 2022), which employed a Variational Autoencoder (VAE) (Kingma, Welling et al. 2019). A two-stage approach was introduced by (Guo et al. 2022a), featuring text2length for sampling motion lengths and a temporal VAE for synthesizing diverse motions. T2M-GPT (Zhang et al. 2023a) also proposed a two-stage method, uniquely utilizing a VQ-VAE and a GPT-like model. With the advent of multi-modal models, MotionCLIP (Tevet et al. 2022) harnessed the latent space of CLIP (Huang et al. 2019). The diffusion model emerged as a novel tool, with MotionDiffuse (Zhang et al. 2022) and ReMoDiffuse (Zhang et al. 2023b) generating vivid, semantically consistent, and high-fidelity motion sequences. This exciting domain continues to evolve, pushing the boundaries of text-driven motion synthesis. Despite these promising developments in text-to-motion methodologies, it is noteworthy that they are predominantly designed for textual inputs. The community currently lacks a unified framework that can competently cater to varying input modalities.

Music to Motion Methods

Research on music-to-dance generation has traditionally relied on low-level music features such as Mel spectrum, Mel Frequency Cepstral Coefficient (MFCC), or short-time

Fourier transform (STFT) spectrum (Griffin and Lim 1984). However, these fail to capture high-level music features like rhythm, beat, and style, which are crucial for harmonious music-motion relations. Early works employed LSTM-autoencoders to generate 3D dance motion from these features (Tang, Jia, and Mao 2018), but their sensitivity to noise limited their effectiveness (Lee et al. 2019). DanceNet (Huang et al. 2020) introduced an auto-regressive model with dilated convolution and GMM loss, supporting long-term music-to-dance generation. The advent of Transformer-based architectures, such as the approach by (Li et al. 2021) that utilized VQVAE and GPT with cross-conditional attention, and (Siyao et al. 2022) that proposed a Full Attention Cross-modal Transformer (FACT) model, brought about significant improvements. Despite these advancements, the majority of music-to-dance methodologies remain specifically designed for music inputs, thereby lacking a unified framework that can adeptly handle diverse input modalities. In this paper, we propose a broader framework, Everything2Motion, designed for human motion synthesis across diverse inputs.

Method

Overview

Given an input modality x (e.g., music clip or text description), we aim to generate a corresponding 9D motion sequence y. This sequence comprises 3D key points and 6D rotation angles, aligning with the definition in the SMPL (Loper et al. 2023) model. The unified human motion synthesis problem can be formally defined as learning a mapping function F:

$$F(x) = y \tag{1}$$

In the context of Everything2Motion, we focus on unifying two tasks: music-to-dance and text-to-motion. For the former, x is a music clip and y is a dance sequence, and for the latter, x is a text description and y is the corresponding human motion sequence. Our goal is to optimize the function F such that it minimizes the difference between the generated motion sequence and the ground truth motion sequence.

As depicted in Fig. 3, our innovative Everything2Motion framework is primarily composed of three modules, each designed to foster a lightweight, smooth, and diversified generation process: 1) **Input-Output Modality Modulation (IOM) module**: This pioneering module incorporates a specialized input adapter for each distinctive input modality. It bridges the dimensional gap between various modalities, enabling a seamless mapping of multi-modal inputs into a shared latent space. This design promotes the integration and understanding of diverse input data, serving as the foundation of our unified approach. 2) **Query-aware Autoencoder (QA) module**: Equipped with an initialized latent query and output query, this module fuses motion-related guidance information and autoregressive pre-conditions into a unified motion latent representation. By employing multiple cross-attention modules, our model maintains focus on both temporal dynamics and content-related details during the generation process. This results in a highly responsive and detail-conscious representation that enhances the naturalness and contextual relevance of the synthesized motions. 3) **Pre-trained Prior Distillation Autoencoder with Variational Representation (PDA-VR) module**: This module utilizes a pre-trained autoencoder equipped with a variational representation to achieve more nuanced and fluid motion generation. By tapping into the wealth of prior knowledge available in extensive expressive motion datasets, this module ensures that the generated sequences are not only smooth and natural but also infused with a rich variety of expressive elements. Next, we introduce each module in detail.

Input-Output Modality Modulation

We present an innovative Input-Output Modality Modulation (IOM) module, strategically designed with distinct Input and Output adapters. This sophisticated design facilitates the transformation of multi-modal inputs into universally applicable latent features. The unique structure of the IOM module allows for the seamless handling of diverse input modalities, all the while ensuring a consistent format for subsequent processing stages.

Input Adapter. The modalities under consideration in our work encompass both textual descriptions and musical audio. For textual descriptions, we employ a pre-trained CLIP text encoder (Radford et al. 2021) to generate initial token embeddings. We then deploy a Transformer encoder layer to yield word-level features L_t, such that:

$$L_t = E_{clip}(T) \tag{2}$$

where E_{clip} denotes the encoding function of the CLIP text encoder and T represents the textual input. Subsequently, a linear layer is used to map these text features into a fixed-shape latent representation.

As for musical audio, we adopt the data preprocessing approach from (Li et al. 2021). Utilizing the *Librosa* library, we extract acoustic features $A = \{a_1, a_2, ..., a_t\}$, which include MFCC, beat, chroma, and the corresponding dance type D_t. The audio encoder is composed of a multi-layer residual convolutional network and a dance-type embedding layer. Here, the dance type is mapped into one-hot encodings and concatenated into a_i in the feature channels, yielding:

$$L_m = E_{music}(A, D_t) \tag{3}$$

where E_{music} denotes the encoding function of the musical audio.

Finally, the latent features are passed through a bidirectional GRU layer for temporal enhancement. The GRU layer's output, x_l, is the adapted input, computed as:

$$x_l = GRU(L_m) \tag{4}$$

This encapsulates the temporally enhanced features. This well-engineered Input Adapter ensures that diverse modalities are effectively translated into compatible latent features, setting the stage for the subsequent motion synthesis process.

Output Adapter. In addition to the aforementioned modules, our model also features an Output Adapter. Comprising

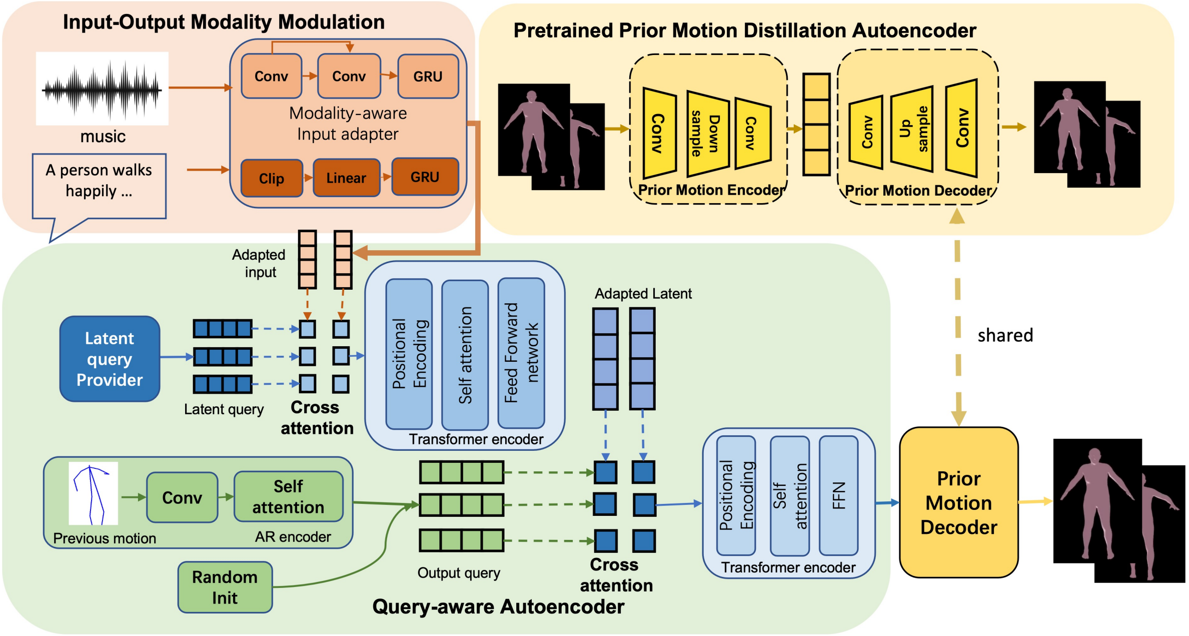

Figure 2: An overview of the unified Everything2Motion framework.

a series of linear blocks, this component is tasked with post-processing the final output, O_{qa}, derived from the Query-aware Autoencoder. The Output Adapter refines this output into a more precise representation O_{out}, suitable for downstream applications. This process can be represented as:

$$O_{out} = f_{linear}(O_{qa}) \tag{5}$$

where f_{linear} denotes the function of the linear blocks in the Output Adapter. This showcases our commitment to producing high-quality, usable outputs, thereby enhancing the utility and robustness of our Everything2Motion framework.

The rationale behind the IOM module's design lies in addressing the challenge of multi-modality inherent in human motion synthesis tasks. Different modalities present varying types of data, each with its unique characteristics and dimensionalities. By introducing separate input and output adapters, we can process these diverse inputs effectively and convert them into a shared latent space. This design ensures the adaptability of our system to a wide range of inputs, thereby enhancing the robustness and flexibility of our Everything2Motion framework.

Query-Aware Autoencoder

In this section, we delve into the architecture and functionality of our proposed Query-aware Autoencoder, a key component of our framework. This Autoencoder interfaces directly with the output from the Input-Output Modality Modulation (IOM) module, taking its processed latent features as input. The primary role of the Query-aware Autoencoder is to further refine these features into a format that is optimally suited for output motion sequence generation. It essentially serves as a bridge, effectively translating the general latent features from the IOM module into a representation that is finely tuned for the downstream task of motion synthesis.

Our Query-aware Autoencoder can be viewed as a reimagined version of the Transformer encoder-decoder structure, with a unique twist. The encoder receives an arbitrarily initialized latent query, q_l, while the decoder input is an output-query, provided by a dynamic provider determined by the generative mode.

$$Q_{encoder} = Encoder(q_l) \tag{6}$$

$$Q_{decoder} = Decoder(Q_{provider}) \tag{7}$$

Latent Query Encoder. In our system design, an input-aware cross-attention mechanism is employed within the latent query. Here, we let $Q, K = q_l$, and $V = x_l$, effectively integrating the latent query with the multi-modal input. This cross-attention mechanism is formulated as:

$$A_l = CrossAttention(q_l, x_l) \tag{8}$$

Upon activation, the latent query is passed through several Transformer encoder layers for further feature extraction and temporal enhancement. These layers comprise self-attention blocks, Fixed Positional Embedding, and a Feed Forward network. Given that our generation task operates in a purely sequence-to-sequence mode, we have eschewed the use of causal masks in the multi-head attention mechanism.

This processing pipeline culminates in the production of an adapted latent query, L_m, which carries a wealth of modality-related information and is output in a consistent shape. This can be represented as:

$$L_m = Transformer(A_l) \tag{9}$$

Output-Query Decoder. Concurrently, an output query O_q is initialized via a modality-related output-query provider. This design provides the flexibility needed to handle both non-autoregressive and autoregressive tasks, each requiring different output-query provider structures.

For tasks such as text2motion, the output query is randomly initialized with learnable parameters conforming to the given output shape. In contrast, for tasks like music2dance, we follow the FACT approach (Li et al. 2021), where dance generation is partially autoregressive, conditioned on previous frames. Consequently, the output-query provider for such tasks is a motion encoder comprising Residual Convolution blocks and a self-attention block, with the previous motion sequence serving as an additional input.

Similar to a Transformer decoder, a cross-attention mechanism is then applied between the adapted latent query L_m and output query O_q. This can be represented as:

$$C_m = CrossAttention(L_m, O_q) \tag{10}$$

The objective of this mechanism is to map the latent motion space and contextual relationship with previous movements into a fixed output shape of $T * D$. Here, T represents the length of the generated motion sequence, and D denotes the dimension of different motion representations (for example, 21 keypoints for KitML, 22 keypoints for AIST++, due to varying keypoint numbers in skeletons).

To further enhance the smoothness and temporal consistency of the output, the adapted output query O_q is passed through the previously discussed Output Adapter. This can be represented as:

$$O_{out} = OutputAdapter(C_m) \tag{11}$$

Building on our previous discussion, the Query-aware Autoencoder proves highly effective for human motion synthesis. Its latent query encoder captures the essential attributes of the input modality through cross-attention, while the output-query decoder maps the adapted latent query into a suitable output shape for generating motion sequences. This results in motion sequences that are contextually coherent and exhibit smooth, temporally consistent movements. Hence, the Query-aware Autoencoder refines features effectively, enhancing the quality of the synthesized human motion sequences.

Prior Motion Distillation Decoder

In the domain of human motion synthesis, our primary goal is to create sequences that are smooth, abide by physical laws, and exhibit diversity. Traditional approaches often rely on modality-specific measures to meet these criteria. However, in our unified framework, devising ad-hoc modules for each modality is unfeasible.

To address this, we propose leveraging the inherent priors found in human motion sequences. More specifically, we train a Motion-based Prior Variational Autoencoder (MP-VAE) to extract these priors and incorporate them into our unified framework.

Our MP-VAE employs a symmetric encoder-decoder structure. The encoder uses convolutional layers and down-sampling operations to derive the latent feature z from the input data:

$$z = Encoder(x) \tag{12}$$

In contrast, the decoder reconstructs the human motion from the latent representation:

$$x' = Decoder(z) \tag{13}$$

The goal of the MP-VAE is to minimize the reconstruction error and align the posterior distribution $q_\theta(z|x)$ as closely as possible to the likelihood $q_\theta(x|z)$. This is encapsulated in the Evidence Lower Bound (ELBO):

$$ELBO = E_{q_\theta(z|x_i)}[\log p_\theta(x_i|z)] - D_{KL}(q_\theta(z|x_i)||p(z)) \tag{14}$$

In our specific implementation, the optimization equation of our VAE is given by:

$$L_{vae} = D_{KL}(z||N(0, I)) + L_2(x_{recon} - x) \tag{15}$$

This equation comprises a reconstruction L2 loss and the Kullback-Leibler (KL) divergence of the variational space generated.

Once the MP-VAE is adequately trained, we distill the human motion priors by directly migrating the pre-trained decoder into our framework, termed Prior Motion Distillation Decoder, as depicted in Fig.3.

Through the use of the Prior Motion Distillation Decoder, we showcase an effective strategy for synthesizing human motion sequences that are smooth, physically plausible, and diverse within a unified framework.

Loss

Given the inherently sequential nature of motion, where each movement is highly dependent on its preceding and succeeding movements, we adopt a velocity-aware loss function to gauge the cross-frame distance errors between the generated motions (g_j) and the ground truth targets (t_j), where j denotes the frame number in a short sequence. This approach, reminiscent of the methodology employed in (Guo et al. 2023), can be expressed as follows:

$$L_{prev} = ||(g_j - g_{j-1}) - (t_j - t_{j-1})||^2 \tag{16}$$

$$L_{latter} = ||(g_{j+1} - g_j) - (t_{j+1} - t_j)||^2 \tag{17}$$

$$L_{loss} = ||g_j - t_j||^2 + \alpha(L_{prev} + L_{latter}) \tag{18}$$

Here, α is a weighting factor that modulates the influence of the velocity terms in the total loss function.

The inclusion of this velocity-continuous loss function substantially enhances the temporal smoothness of the generated motions, particularly when conforming to discrete text descriptions. This refined loss function hence plays an indispensable role in ensuring the synthesis of high-quality, temporally consistent motion sequences.

Method	FID ↓	R precision (Top 1) ↑	R precision (Top 2) ↑	R precision (Top 3) ↑	Diversity ↑	MM-Dist ↓	Modality
Ground Truth	0.031	0.424	0.649	0.779	11.08	2.788	–
Test2Gesture	12.12	0.156	0.255	0.338	9.334	6.964	–
Language2Pose	6.545	0.221	0.373	0.483	9.037	5.147	–
MoCoGAN	94.41	0.037	0.072	0.106	0.462	10.40	0.250
Guo et.al	2.770	0.370	0.569	0.693	10.91	3.401	1.482
Motiondiffuse	1.954	0.417	0.621	0.739	11.10	2.958	0.730
E2M (Ours)	1.060	0.385	0.574	0.685	11.15	3.457	0.051

Table 1: Performance on Text2Motion task on KitML dataset.

Experiments

Datasets

In our study, we utilize two key datasets: KIT Motion-Language (KIT-ML) for text-to-motion (text2motion) tasks, and AIST++ for music-to-dance (music2dance) tasks. The KIT-ML dataset (Mandery et al. 2015) includes 3911 motion sequences paired with 1 to 4 English language descriptions, spanning diverse motion categories such as locomotion, manipulation, and sports. It is derived from the KIT (Plappert, Mandery, and Asfour 2016) and CMU (Merel et al. 2017) databases. For music2dance tasks, we use the AIST++ dataset (Tsuchida et al. 2019), the largest public 3D human dance dataset. It contains 1408 sequences represented as joint rotations and root trajectories from 30 subjects across ten dance genres. Each sequence's data includes 9 camera views, 2D and 3D human joint locations in the COCO-format (Lin et al. 2014), and 24 SMPL (Loper et al. 2015) pose parameters. These datasets, each tailored to a specific task of text2motion or music2dance, together provide a comprehensive foundation for exploring unified human motion synthesis.

Implementation Detail

Using SMPL-X parameter models, our motion format handles KitML and AIST++ keypoints, with 21 and 24 joints respectively. A consistent preprocessing procedure calculates six additional rotation angles and introduces the corresponding transition matrix. Music processing, following (Tsuchida et al. 2019), extracts MFCC, beat, peak, and Chroma features at 25 fps per sequence. Text2motion's input adapter combines a pre-trained Clip encoder, a 2-layer transformer encoder, and a single linear layer. Conversely, the music2motion adapter uses convolutional layers, a 5-D type embedding network output, and a double-layer transformer encoder. Both share a linear-based 3-layer network output adapter. Text2motion's Query-aware Autoencoder has 128 latent codes and features a 128 feature-shaped autoencoder with four self-attention and two cross-attention layers. Music2dance's larger latent space has 256 latent codes, a 128 feature-shaped autoencoder, and two self-attention layers. The output-query provider for music2motion uses a 3-layer convolutional network. The prior distillation autoencoder, symmetric with 3-layer convolutional blocks, has a 256 feature-shaped bottleneck. This autoencoder is pre-trained and jointly trained with the Everything2Motion framework.

Method	FID ↓	Diversity ↑	Beat Align ↑
Ground Truth	10.60	7.45	0.237
Li et.al	43.46	3.32	0.160
Dancenet	25.49	2.85	0.143
DanceRevolution	25.92	4.87	0.195
FACT	22.11	6.18	0.221
E2M (Ours)	38.23	6.68	0.248

Table 2: Performance on Music2Dance task on AIST++ dataset.

Evaluation Metrics

Our methodology is evaluated employing distinct metrics for the two tasks: text-to-motion and music-to-motion. Specifically, we apply R Precision, Frechet Inception Distance (FID), Diversity, Multimodal Distance (MM-Dist), and Multimodality (MModality) for Text2Motion evaluation. R Precision assesses the Euclidean distances between motion and text embeddings. FID gauges the distribution distance between features of generated and real motions. Diversity measures the variation in motion sequences created from test set descriptions. MM-Dist calculates the difference between text features from the description and generated motions. MModality measures differences in joint positions in motion sequences generated from a single text description. We use FID to measure the overall quality of the generated dance movements, Diversity to compute the average Euclidean distance across generated motions and Beat-align to measure the synchronization for motion and music on the AIST++ test set for Music2Dance evaluation, calculated primarily in the geometry feature space for a simplified evaluation.

Results

In this section, we show the quantitative results of our method and compare it with strong modality specific baselines. For qualitative results, please refer to the supplementary material.

Performance on Text2Motion task: Our E2M (Everything2Motion) model, as demonstrated in Table 1, exhibits exceptional performance across key evaluation metrics in the Text2Motion task, outperforming other modality-specific methods. A defining characteristic of our model is its unified framework, capable of handling various modalities of input data, thereby providing an innovative solution in the field of motion generation. In particular, the E2M model's strikingly low FID score of 1.060, compared to the 2.770 of Guo et al. (Guo et al. 2022a) and 1.954 of Motiondiffuse (Zhang et al. 2022), is a testament to its excellent ability to generate high-quality and realistic motions. The FID score measures the distance between the generated motion and the actual motion, and a lower score indicates a closer match to reality. E2M's low score reflects its exceptional capacity to generate outputs that closely mirror real-world motions. Moreover, E2M shines in terms of R precision, another critical metric in motion generation. It scores 0.385, 0.574, and 0.685 at Top 1, Top 2, and Top 3 levels, respectively. These high scores demonstrate E2M's proficiency in effectively capturing and integrating text-related information into motion synthesis. The ability to generate motion that correctly corresponds to the provided textual description is a vital aspect of the Text2Motion task, and E2M excels in this regard. E2M also impresses with its Diversity score of 11.15, which is very close to the 11.08 score of the ground truth motion. This high score underscores E2M's ability to generate a diverse range of motion outputs. It's a significant accomplishment, as modality-specific methods often struggle to balance diversity and accuracy. In conclusion, our E2M model marks a significant advancement in the Text2Motion task.

Performance on Music2Dance task: In the arena of music-to-motion tasks, our Everything2Motion (E2M) framework has been rigorously benchmarked against the FACT (Siyao et al. 2022), Li et al. (Li et al. 2021), and other SOTA methods, using the AIST++ dataset as the evaluative standard. As evidenced in Table 2, our framework excels particularly in the Beat-Align metric, which measures the synchronization between the given music and dance movements. Achieving a score of 0.248, E2M surpasses all other established methods in this category, demonstrating its superior ability to generate dance sequences that harmonize rhythmically with the music. Despite the complex nature of this task, our model also achieves comparable results in the other metrics. The FID score of 38.23 and the diversity score of 6.68, while not the highest, are competitive when compared with other state-of-the-art models. These results underscore the robustness of our model across different aspects of dance generation. It is important to emphasize that our E2M framework is designed as a unified model, with the aim of achieving solid performance across various modality inputs. The pursuit of this work is to refine the model's ability to generalize across different modalities. Despite this broad focus, our method achieves results on par with existing state-of-the-art modality-specific approaches in the music-to-motion task. This speaks volumes about the versatility and superiority of our approach.

Running Time Analysis: Our Everything2Motion framework is also characterized by its lightweight design, which enables real-time implementation. To illustrate this point, we conduct an inference time test for our method and compared it with the state-of-the-art approaches, as demonstrated in Table 3. For the task of text-to-motion, our method is able to generate 120 frames of motion in just 0.0904s on a 3090 gpu. In contrast, the method proposed by Guo et al. (Guo et al. 2022a) requires 0.622s to achieve the same task. This benchmark illustrates that our method offers an order of magnitude improvement in terms of inference speed. Similarly, for the music-to-dance task, our method predicts 240 frames of motion in a mere 0.0384s, a speed that far surpasses the FACT method (Siyao et al. 2022). These results unequivocally demonstrate the efficiency of our E2M framework. Its superior speed in generating motion frames from both text and music inputs highlights its potential in real-time applications, marking it out as a significant advancement in the field.

Method	Task	Time (s) ↓
Guo et al.	Text2Motion	0.6220
E2M	Text2Motion	**0.0904**
FACT	Music2Dance	0.1082
E2M	Music2Dance	**0.0384**

Table 3: Inference time comparison for Text2Motion and Music2Dance tasks.

Visualization: In order to provide a more comprehensive evaluation of the quality of our generated frames, we showcase a selection of generated results derived from the KitML dataset for text2motion tasks in Fig. 3. The results encompass three distinct movements, namely: 1) Being pushed 2) Walking in degree angles 3) Stretching arms. This delineation of results serves to demonstrate the versatility and capability of our model in effectively translating varied textual descriptions into corresponding motion sequences.

Ablation Study

Ablation study on structure design: To validate the effectiveness of key model structure designs, we conducted an ablation study on the text-to-motion task using the KitML dataset. The study primarily focused on the pre-trained Clip encoder and the Prior Motion Distillation Decoder (PMD). As illustrated in Table 4, when the pre-trained Clip encoder is not utilized (E2M w/o Clip), the model's ability to generate matching motion sequences during input adapter extraction is substantially compromised, leading to a higher FID score of 4.624 and a lower R precision top1 score of 0.287. This indicates the crucial role of the Clip encoder in the model's structure. Similarly, when we bypass the Prior Motion Distillation Decoder (E2M w/o PMD), it results in a further increase in the FID score to 4.258 and a drop in the R precision top1 score to 0.244, indicating a loss of temporal consistency in the generated motion sequences due to an unstable latent space. In contrast, the fully equipped E2M model, which includes both the Clip encoder and PMD, achieves a significantly lower FID score of 1.060 and a higher R precision top1 score of 0.385.

Ablation study on query-aware autoencoder: The selection of latent shape in the Query-Aware Autoencoder is

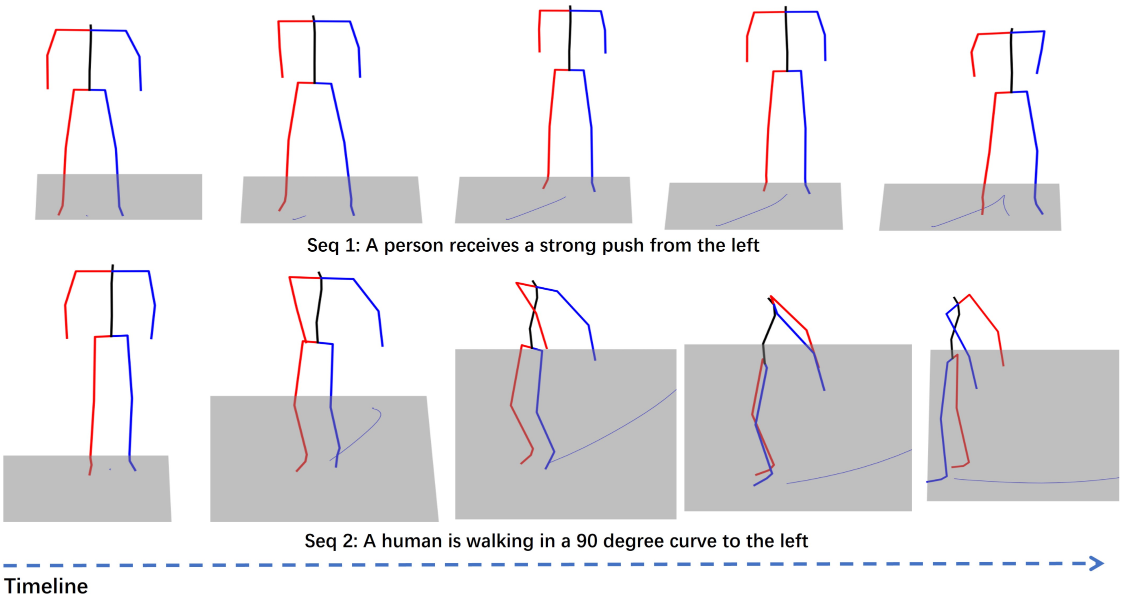
Seq 1: A person receives a strong push from the left
Seq 2: A human is walking in a 90 degree curve to the left
Timeline

Figure 3: Visualization of generated results

Method	FID ↓	R precision top1 ↑
GT	0.031	0.424
E2M w/o PMD	4.258	0.244
E2M w/o Clip	4.624	0.287
E2M	**1.060**	**0.385**

Table 4: Ablation study on KitML dataset for Everything2Motiong's framework architecture.

			KitML			
Method	Num	Dim	FID ↓	Top1	Top2	Top3
GT	-	-	0.031	0.424	0.649	0.779
E2M	64	128	2.086	0.277	0.462	0.577
E2M	128	64	2.468	0.361	0.544	0.656
E2M	128	256	1.243	0.274	0.472	0.607
E2M	256	256	1.147	0.303	0.472	0.580
E2M	128	128	**1.160**	**0.385**	**0.574**	**0.685**

Table 5: Ablation study results of the E2M framework structure. We evaluate the performance difference for the model without Pre-trained Prior Motion Decoder and Clip encoder.

found to significantly influence the overall performance of the model. An unsuitable latent query can lead to unstable training and mode collapse, underscoring the importance of an appropriate latent shape selection. To systematically evaluate this, we conduct an extensive ablation study focusing on the selection of the latent shape. Specifically, we assess the impact of varying the number of latent codes and latent channels, evaluating them at levels of 64, 128, and 256. As indicated in Table 5, enlarging either the number of codes or channels tends to cause a slight degradation of the generative performance across all metrics. The Frechet Inception Distance (FID) score is particularly impacted, which could be attributed to the overabundance of information present in the latent query due to an unsuitable latent shape. This excessive information makes it more challenging for the output query to extract useful information during the training procedure. Conversely, our ablation experiments suggest that a higher number of latent codes and shape dimensions enable the model to generate more natural and plausible results, which are characterized by lower FID scores. This underscores the need for a judicious balance in the selection of latent shape parameters to ensure optimal model performance.

Conclusion

In this paper, we have introduced Everything2Motion, a novel unified model framework that aims to revolutionize human motion synthesis in the dynamic fields of film and game development. By integrating the Input-Output Modality Modulation module, the Query-aware Autoencoder, and the Prior Motion Distillation Decoder, Everything2Motion eliminates the need for modality-specific frameworks, facilitates efficient latent motion generation, and enhances the naturalness and fluidity of the final skeleton sequences. Our comprehensive experiments conducted on several public datasets have demonstrated the effectiveness of Everything2Motion, providing compelling evidence of its potential for practical implementation. While Everything2Motion presents a promising approach for diverse motion synthesis tasks, there remain areas for exploration. Future work may extend its application to other modalities and datasets, such as video2motion or image2motion.

References

Ahuja, C.; and Morency, L.-P. 2019. Language2pose: Natural language grounded pose forecasting. In *2019 International Conference on 3D Vision (3DV)*, 719–728. IEEE.

Barsoum, E.; Kender, J.; and Liu, Z. 2018. Hp-gan: Probabilistic 3d human motion prediction via gan. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 1418–1427.

Dabral, R.; Mughal, M. H.; Golyanik, V.; and Theobalt, C. 2023. Mofusion: A framework for denoising-diffusion-based motion synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9760–9770.

Fan, Z.; Li, F.; Liu, H.; He, J.; and Du, X. 2022. Human Pose Driven Object Effects Recommendation. *arXiv preprint arXiv:2209.08353*.

Griffin, D.; and Lim, J. 1984. Signal estimation from modified short-time Fourier transform. *IEEE Transactions on acoustics, speech, and signal processing*, 32(2): 236–243.

Guo, C.; Zou, S.; Zuo, X.; Wang, S.; Ji, W.; Li, X.; and Cheng, L. 2022a. Generating diverse and natural 3d human motions from text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5152–5161.

Guo, C.; Zuo, X.; Wang, S.; and Cheng, L. 2022b. Tm2t: Stochastic and tokenized modeling for the reciprocal generation of 3d human motions and texts. In *European Conference on Computer Vision*, 580–597. Springer.

Guo, W.; Du, Y.; Shen, X.; Lepetit, V.; Alameda-Pineda, X.; and Moreno-Noguer, F. 2023. Back to mlp: A simple baseline for human motion prediction. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 4809–4819.

Holden, D.; Saito, J.; and Komura, T. 2016. A deep learning framework for character motion synthesis and editing. *ACM Transactions on Graphics (TOG)*, 35(4): 1–11.

Huang, C.; Lin, C.-E.; Yang, Z.; Kong, Y.; Chen, P.; Yang, X.; and Cheng, K.-T. 2019. Learning to film from professional human motion videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4244–4253.

Huang, R.; Hu, H.; Wu, W.; Sawada, K.; Zhang, M.; and Jiang, D. 2020. Dance revolution: Long-term dance generation with music via curriculum learning. *arXiv preprint arXiv:2006.06119*.

Jaegle, A.; Borgeaud, S.; Alayrac, J.-B.; Doersch, C.; Ionescu, C.; Ding, D.; Koppula, S.; Zoran, D.; Brock, A.; Shelhamer, E.; et al. 2021. Perceiver io: A general architecture for structured inputs & outputs. *arXiv preprint arXiv:2107.14795*.

Khatib, O.; Demircan, E.; De Sapio, V.; Sentis, L.; Besier, T.; and Delp, S. 2009. Robotics-based synthesis of human motion. *Journal of physiology-Paris*, 103(3-5): 211–219.

Kingma, D. P.; Welling, M.; et al. 2019. An introduction to variational autoencoders. *Foundations and Trends® in Machine Learning*, 12(4): 307–392.

Lee, H.-Y.; Yang, X.; Liu, M.-Y.; Wang, T.-C.; Lu, Y.-D.; Yang, M.-H.; and Kautz, J. 2019. Dancing to music. *Advances in neural information processing systems*, 32.

Lee, J.; Chai, J.; Reitsma, P. S.; Hodgins, J. K.; and Pollard, N. S. 2002. Interactive control of avatars animated with human motion data. In *Proceedings of the 29th annual conference on Computer graphics and interactive techniques*, 491–500.

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

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, 740–755. Springer.

Liu, W.; Piao, Z.; Min, J.; Luo, W.; Ma, L.; and Gao, S. 2019. Liquid warping gan: A unified framework for human motion imitation, appearance transfer and novel view synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 5904–5913.

Loper, M.; Mahmood, N.; Romero, J.; Pons-Moll, G.; and Black, M. J. 2015. SMPL: A Skinned Multi-Person Linear Model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6): 248:1–248:16.

Loper, M.; Mahmood, N.; Romero, J.; Pons-Moll, G.; and Black, M. J. 2023. SMPL: A skinned multi-person linear model. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, 851–866.

Mandery, C.; Terlemez, Ö.; Do, M.; Vahrenkamp, N.; and Asfour, T. 2015. The KIT whole-body human motion database. In *2015 International Conference on Advanced Robotics (ICAR)*, 329–336. IEEE.

Menache, A. 2000. *Understanding motion capture for computer animation and video games*. Morgan kaufmann.

Merel, J.; Tassa, Y.; TB, D.; Srinivasan, S.; Lemmon, J.; Wang, Z.; Wayne, G.; and Heess, N. 2017. Learning human behaviors from motion capture by adversarial imitation. *arXiv preprint arXiv:1707.02201*.

Petrovich, M.; Black, M. J.; and Varol, G. 2021. Action-conditioned 3D human motion synthesis with transformer VAE. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10985–10995.

Petrovich, M.; Black, M. J.; and Varol, G. 2022. TEMOS: Generating diverse human motions from textual descriptions. In *European Conference on Computer Vision*, 480–497. Springer.

Plappert, M.; Mandery, C.; and Asfour, T. 2016. The KIT motion-language dataset. *Big data*, 4(4): 236–252.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Siyao, L.; Yu, W.; Gu, T.; Lin, C.; Wang, Q.; Qian, C.; Loy, C. C.; and Liu, Z. 2022. Bailando: 3d dance generation by actor-critic gpt with choreographic memory. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11050–11059.

Sun, G.; Wong, Y.; Cheng, Z.; Kankanhalli, M. S.; Geng, W.; and Li, X. 2020. DeepDance: music-to-dance motion choreography with adversarial learning. *IEEE Transactions on Multimedia*, 23: 497–509.

Tang, T.; Jia, J.; and Mao, H. 2018. Dance with melody: An lstm-autoencoder approach to music-oriented dance synthesis. In *Proceedings of the 26th ACM international conference on Multimedia*, 1598–1606.

Tevet, G.; Gordon, B.; Hertz, A.; Bermano, A. H.; and Cohen-Or, D. 2022. Motionclip: Exposing human motion generation to clip space. In *European Conference on Computer Vision*, 358–374. Springer.

Tsuchida, S.; Fukayama, S.; Hamasaki, M.; and Goto, M. 2019. AIST Dance Video Database: Multi-Genre, Multi-Dancer, and Multi-Camera Database for Dance Information Processing. In *ISMIR*, volume 1, 6.

Yan, X.; Rastogi, A.; Villegas, R.; Sunkavalli, K.; Shechtman, E.; Hadap, S.; Yumer, E.; and Lee, H. 2018. Mt-vae: Learning motion transformations to generate multimodal human dynamics. In *Proceedings of the European conference on computer vision (ECCV)*, 265–281.

Zhang, J.; Zhang, Y.; Cun, X.; Huang, S.; Zhang, Y.; Zhao, H.; Lu, H.; and Shen, X. 2023a. T2m-gpt: Generating human motion from textual descriptions with discrete representations. *arXiv preprint arXiv:2301.06052*.

Zhang, M.; Cai, Z.; Pan, L.; Hong, F.; Guo, X.; Yang, L.; and Liu, Z. 2022. Motiondiffuse: Text-driven human motion generation with diffusion model. *arXiv preprint arXiv:2208.15001*.

Zhang, M.; Guo, X.; Pan, L.; Cai, Z.; Hong, F.; Li, H.; Yang, L.; and Liu, Z. 2023b. ReMoDiffuse: Retrieval-Augmented Motion Diffusion Model. *arXiv preprint arXiv:2304.01116*.

Zhou, Z.; and Wang, B. 2023. Ude: A unified driving engine for human motion generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5632–5641.

Zhuang, W.; Wang, C.; Chai, J.; Wang, Y.; Shao, M.; and Xia, S. 2022. Music2dance: Dancenet for music-driven dance generation. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 18(2): 1–21.