Language-guided Human Motion Synthesis with Atomic Actions

Yuanhao Zhai
University at Buffalo
Buffalo, NY, USA
yzhai6@buffalo.edu

Mingzhen Huang
University at Buffalo
Buffalo, NY, USA
mhuang33@buffalo.edu

Tianyu Luan
University at Buffalo
Buffalo, NY, USA
tianyulu@buffalo.edu

Lu Dong
University at Buffalo
Buffalo, NY, USA
ludong@buffalo.edu

Ifeoma Nwogu
University at Buffalo
Buffalo, NY, USA
inwogu@buffalo.edu

Siwei Lyu
University at Buffalo
Buffalo, NY, USA
siweilyu@buffalo.edu

David Doermann
University at Buffalo
Buffalo, NY, USA
doermann@buffalo.edu

Junsong Yuan
University at Buffalo
Buffalo, NY, USA
jsyuan@buffalo.edu

ABSTRACT

Language-guided human motion synthesis has been a challenging task due to the inherent complexity and diversity of human behaviors. Previous methods face limitations in generalization to novel actions, often resulting in unrealistic or incoherent motion sequences. In this paper, we propose ATOM (ATomic mOtion Modeling) to mitigate this problem, by decomposing actions into atomic actions, and employing a curriculum learning strategy to learn atomic action composition. First, we disentangle complex human motions into a set of atomic actions during learning, and then assemble novel actions using the learned atomic actions, which offers better adaptability to new actions. Moreover, we introduce a curriculum learning training strategy that leverages masked motion modeling with a gradual increase in the mask ratio, and thus facilitates atomic action assembly. This approach mitigates the overfitting problem commonly encountered in previous methods while enforcing the model to learn better motion representations. We demonstrate the effectiveness of ATOM through extensive experiments, including text-to-motion and action-to-motion synthesis tasks. We further illustrate its superiority in synthesizing plausible and coherent text-guided human motion sequences.

CCS CONCEPTS

• **Computing methodologies** → **Activity recognition and understanding**.

KEYWORDS

language-guided human motion synthesis, atomic action, masked motion modeling, curriculum learning

MM '23, October 29-November 3, 2023, Ottawa, ON, Canada

ACM ISBN 979-8-4007-0108-5/23/10…$15.00
https://doi.org/10.1145/3581783.3612289

ACM Reference Format:
Yuanhao Zhai, Mingzhen Huang, Tianyu Luan, Lu Dong, Ifeoma Nwogu, Siwei Lyu, David Doermann, and Junsong Yuan. 2023. Language-guided Human Motion Synthesis with Atomic Actions. In *Proceedings of the 31st ACM International Conference on Multimedia (MM '23), October 29-November 3, 2023, Ottawa, ON, Canada.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3581783.3612289

1 INTRODUCTION

Language-guided human motion synthesis is a critical and challenging task, with widespread applications in virtual reality, video games, animation, and human-computer interaction. The ability to generate realistic and diverse human motions based on textual descriptions can enable more intuitive control over virtual characters, as well as seamless integration of user-generated content in various multimedia systems. Despite previous conditional action generation models [5, 8, 9, 20, 34, 47, 51, 53, 54, 58] have exploited leveraging closed-set action labels to synthesize human motion, few of them can work beyond the training action labels, not to mention open language descriptions.

In recent years, various methods have been proposed for the language-guided human motion synthesis task [2, 7, 10, 12–14, 17, 30, 35, 42, 43, 48, 57]. While these methods have demonstrated considerable progress in generating human motions, they may suffer from synthesizing discontinuities and unrealistic motion transitions when dealing with actions that are not well-represented in the dataset. As these methods rely heavily on the availability and diversity of training data, they may struggle to generate plausible motion sequences for rare or unseen actions, leading to abrupt transitions and incoherent movement patterns. Furthermore, when synthesizing complex sequential motion behaviors, these methods often struggle with capturing dependencies in motion sequences, which is crucial for ensuring smooth and natural motion transitions between different types of actions, as illustrated in Fig. 1.

We propose ATOM (ATomic mOtion Modeling), a novel approach for language-guided human motion synthesis that effectively addresses the limitations of previous methods. First, ATOM decomposes actions into atomic components, enabling the generation of diverse and coherent motion sequences by assembling the learned

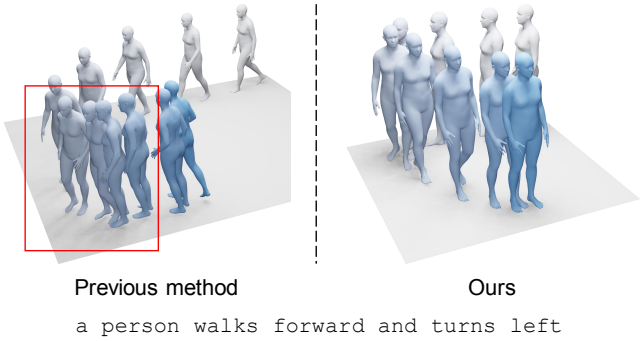


Figure 1: Motion synthesis results comparison. Previous T2M [13] generates an unrealistic motion transition between "`walk forward`" and "`turn left`" (see motion inside the red box), while our ATOM generates coherent motion. The color saturation increases as the motion progresses.

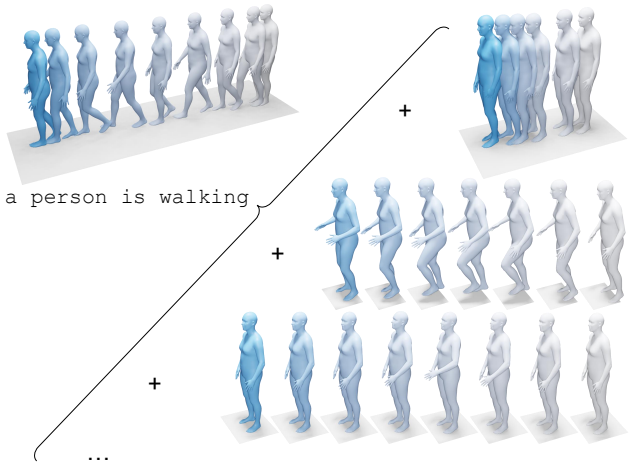


Figure 2: Illustration of atomic action. The action "`a person is walking`" can be decomposed into a set of atomic actions: from top to bottom, whole body translation, lifting leg, hand movements and so on.

atomic actions. Additionally, by employing a masked motion modeling curriculum learning strategy, our method learns more expressive motion representations and effectively captures long-range dependencies in motion sequences.

In our proposed ATOM, we utilize a transformer-based conditional variational autoencoder (CVAE) framework [46] to achieve atomic action decomposition and assembly, which provides a more expressive and flexible representation of human motions. The atomic action codebook, designed as a set of learnable feature vectors, serves as the key and value for the cross-attention module in the Transformer decoder. These atomic actions, learned in an end-to-end manner, can effectively represent a wide range of short-term basic human movements, such as raising hands and lifting legs. Fig. 2 showcases an example, where the "`walking`" action can be decomposed into a set of atomic actions. To ensure the atomicity and effectiveness of the learned actions, we apply two additional constraints: diversity and sparsity selection constraints. The diversity constraint encourages the atomic actions to be distinct from each other, allowing for a richer and more versatile action representation. The sparsity selection constraint enforces atomic actions to be compact and focused during composition, ensuring that the learned atomic actions retain their atomicity. This constraint aids in the efficient reconstruction of complex actions by combining a minimal set of atomic actions while preserving their individual characteristics. The combination of the Transformer-based architecture and these constraints allows our ATOM to efficiently decompose and assemble complex actions for improved human motion synthesis, ultimately generating more diverse, coherent and realistic motion sequences based on language input.

Alongside the atomic action decomposition, we incorporate a curriculum learning strategy into our ATOM to further improve the robustness of generating coherent and diverse motion sequences while capturing long-range dependencies. This strategy is based on masked motion modeling and involves gradually increasing the mask ratio during the training process. Such a progressive learning approach provides the model with an opportunity to learn simpler patterns and dependencies in the early stages of training, while incrementally introducing more complex and challenging aspects of the motion sequences as training progresses. This approach not only mitigates potential convergence issue, but also helps the model build a more robust and generalizable understanding of motion data, encompassing both local and long-range motion patterns. As a result, ATOM is better equipped to synthesize realistic and diverse human motions based on language inputs, even when dealing with rare or unseen action labels/descriptions.

To summarize, our contributions are as follows:

- We introduce a transformer-based CVAE framework that decomposes complex actions into a set of atomic actions, enabling more effective representation and manipulation of motion sequences. This approach allows for generating diverse and coherent motion sequences, even for rare or unseen action labels/descriptions.
- We incorporate a curriculum learning strategy based on masked motion modeling, which gradually increases the mask ratio during training. This strategy enables ATOM to better capture dependencies in motion sequences, ensuring smooth and natural motion transitions between different actions.
- We provide a comprehensive evaluation of ATOM, demonstrating its effectiveness in generating coherent and diverse motion sequences. Our method significantly improves over previous approaches in the text-to-motion and action-to-motion tasks.

2 RELATED WORK

Human motion synthesis Directly synthesizing human motion has always been a challenging yet ideal task in computer vision and graphics. As the field has progressed, methods have evolved from generating skeleton-based motion synthesis [5, 9, 20, 53, 54] to more realistic SMPL motion synthesis [8, 34, 47, 51], bringing us closer to generating authentic human motions in real-world scenarios [28, 29, 56]. Tilmanne *et al.* [44] employed Gaussian distributions to model the variability of walk cycles for each emotion and the length of each cycle. Zhang *et al.*[58] further explored generating unbounded human motion through a cross-conditional, two-stream

variational RNN architecture. CSGN [52] jointly modeled structures in temporal and spatial dimensions, allowing bidirectional transforms between the latent and observed spaces to handle semantic manipulation of action sequences. SA-GCNs [54] proposed a variant of GCNs that leverages the self-attention mechanism to adaptively sparsify a complete action graph in the temporal space. ACTOR [34] learned an action-aware latent representation for human motions by training a transformer-based VAE. ActFormer [34] focused on multi-person interactive actions, combining the solid spatio-temporal representation capacity of the Transformer, the generative modeling superiority of GANs, and the inherent temporal correlations from latent prior. INR [6] employs variational implicit neural representations to generate variable-length sequences. Recently, a growing body of work has emerged, focusing on leveraging natural language to synthesize human motions [1, 2, 7, 10, 12–14, 17, 30, 35, 42, 43, 48, 57]. Specifically, Language2Pose [1] learns a joint embedding of language and pose decoder to generate pose sequences. Text2gesture [3] exploits relevant biomechanical features for body expression to create emotive body gestures. Hier [10] introduces a self-supervised method for generating long-range behaviors. T2M [13] employs a curated language encoder to learn crucial words and a duration estimator to synthesize human motions of varying durations.

Language-guided generation Language-guided generation establishes a connection between visual representation and semantic space, enabling more precise control and increased creative possibilities. Synthesizing motion from language, especially when dealing with multiple, varied actions is a challenging task compared to generating images with specific action labels. Numerous prior language-guided generation methods have focused on image generation [23, 24, 26, 32, 50, 55]. Reed *et al.* [39] proposed using a generative adversarial network [11] (GAN) conditioned on text embeddings for image synthesis. DALL·E [38] employed a discrete variational autoencoder (dVAE) to generate diverse images based on text embeddings from GPT-3 [4]. The recently proposed CLIP [37] jointly learns a multi-modal vision-language embedding space with impressive capabilities. Leveraging the power of CLIP [37], StyleCLIP [33] extends StyleGAN [19] to a language-driven generation model using their proposed CLIP-guided mapper. Meanwhile, HairCLIP [49] generates manipulated images from given CLIP [37] text embeddings, further demonstrating the potential of language-driven generation.

3 METHOD

3.1 Problem Formulation

The task of language-guided human motion synthesis aims to generate motion sequences that accurately represent the given textual description. During the training phase, the inputs consist of label-motion pairs $\{(y_i, \boldsymbol{M}_i)\}$, where the label can be either a natural language description (text-to-motion) or a discrete action class (action-to-motion). The motion representation $\boldsymbol{M}_i = [\boldsymbol{p}_1, \ldots, \boldsymbol{p}_T]$ is a sequence of human body representations with length T, where $\boldsymbol{p}_t$ denotes the human body representation at time t. The human body representation can adopt various forms, such as 3D joint locations or SMPL parameters [27]. During the testing phase, the objective is to synthesize motion sequences based on the input language description or action class.

3.2 Conditional Transformer VAE

We utilize a CVAE-based framework, as shown in Fig. 3, which aligns motion representations with categorical conditions. Our ATOM consists of an encoder and a decoder, implemented using Transformer encoder and decoder [46]. The encoder captures the input motion sequence's underlying structure, transforming it into a compact latent representation. The decoder uses the latent representation and text embedding to generate a realistic human motion sequence corresponding to the given condition.

Encoder The encoder accepts the conditional embedding of the label y and the motion representation sequence $[\boldsymbol{p}_t]$, and computes the Gaussian distribution parameters μ and Σ for the motion latent space. A latent variable $\boldsymbol{z}$ is then sampled from $N(\mu, \Sigma)$ using the reparameterization trick [21]. We prepend two tokens, μ_0 and Σ_0, to the input, allowing their corresponding outputs to be regarded as the Gaussian distribution parameters. These two input tokens, μ_0 and Σ_0, are derived from the input embedding through a three-layer multilayer perceptron (MLP). To ensure compatibility, all motion representations $\boldsymbol{p}_t$ are transformed to the same dimension as μ_0 and Σ_0 using a linear layer before entering the Transformer encoder.

Decoder Given a latent vector $\boldsymbol{z}$, we first add a conditional bias to it to incorporate the categorical information. This bias is action-specific and learned from the input embedding through a three-layer MLP. Subsequently, the sum is repeated T times in the temporal dimension, and sinusoidal positional encoding [46] is added as the query input to the Transformer decoder. A set of learnable atomic actions is provided to the Transformer decoder as key and value inputs, enabling the query to be reconstructed using the atomic actions. The output of the decoder is the reconstructed motion representation $\hat{\boldsymbol{M}} = [\hat{\boldsymbol{p}}_t]$, where $\hat{\boldsymbol{p}}_t$ the predicted body representation at time t. It is worth noting that the decoder generates the entire motion sequence in one shot, as opposed to the autoregressive approach used in previous works [34].

Learning Objectives The learning objectives of the CVAE consist of two components: a reconstruction loss $\mathcal{L}_{\text{rec}}$ and a Kullback-Leibler (KL) divergence loss $\mathcal{L}_{\text{KL}}$. The reconstruction loss is employed to minimize the discrepancy between the original motion representation $\boldsymbol{M}$ and the reconstructed motion representation $\hat{\boldsymbol{M}}$ through a mean square error loss:

$$\mathcal{L}_{\text{rec}} = \frac{1}{T}\sum_{t=1}^{T} \|\boldsymbol{p}_t - \hat{\boldsymbol{p}}_t\|_2^2 . \tag{1}$$

On the other hand, the KL divergence loss $\mathcal{L}_{\text{KL}}$ minimizes the distribution difference between the estimated posterior $N(\mu, \Sigma)$ and the prior normal distribution $N(0, I)$. Thus, the CVAE training loss $\mathcal{L}_{\text{CVAE}}$ is a weighted sum of the two terms:

$$\mathcal{L}_{\text{CVAE}} = \mathcal{L}_{\text{rec}} + w_{\text{KL}}\mathcal{L}_{\text{KL}}, \tag{2}$$

where w_{KL} is a weighting hyperparameter.

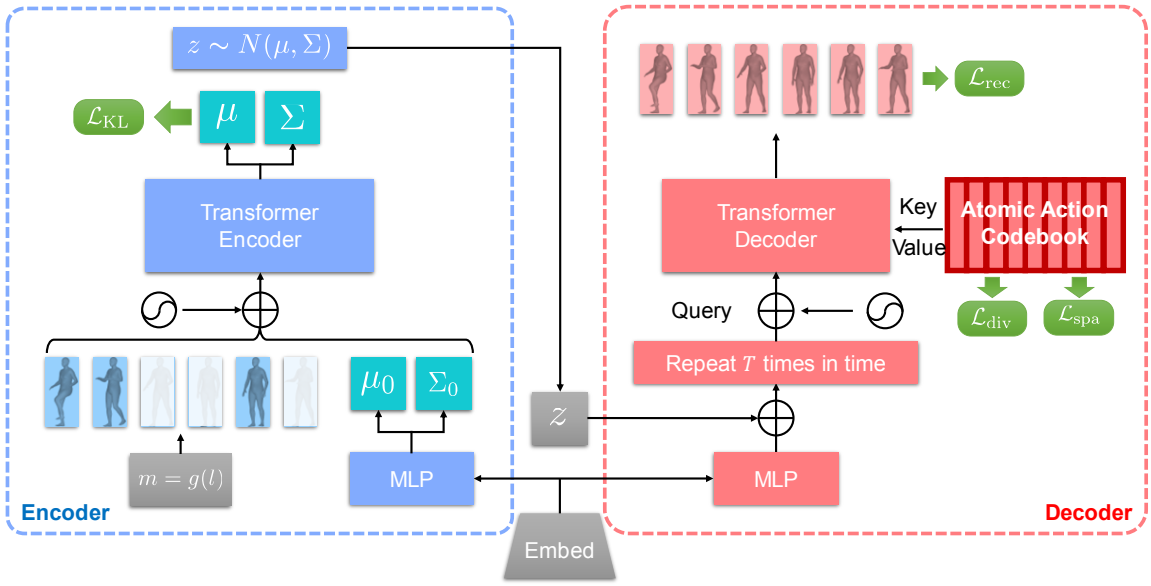


Figure 3: Framework overview. Our ATOM is composed of an encoder and a decoder. The encoder processes text embedding of the action label and masked motion sequence, outputting a latent vector z. The decoder receives this latent vector z along with the text embedding. A set of learnable atomic actions, referred to as the atomic action codebook, is fed into the decoder as key and value for the cross-attention layers. As a result, the generated motion sequences are assembled from the atomic actions.

3.3 Atomic Action Codebook

The motivation for using atomic actions in our method stems from the observation that human motions, despite their apparent complexity, can often be decomposed into more specific, repetitive, and atomic elements. By breaking down complex actions into a series of atomic actions, our model can more effectively learn the underlying structure of human motion and capture the relationships between different actions. Furthermore, this decomposition facilitates the generation of diverse and realistic motion sequences, as well as the synthesis of novel actions by recombining the learned atomic elements. It also enables our method to better generalize across different action classes and leverage the power of textual descriptions in guiding the synthesis process.

Building on the idea of atomic actions, we introduce a learnable atomic action codebook, which serves as a basis for representing and generating complex human motions within our Transformer-based architecture. This codebook consists of a collection of atomic actions that can be combined and assembled in various ways to produce a diverse range of motion sequences. Formally, the codebook is implemented as a learnable matrix $\boldsymbol{A} \in \mathbb{R}^{N \times D}$, where N represents the number of atomic actions, and D denotes the hidden dimension. Each row of the matrix corresponds to an atomic action, capturing its unique characteristics in the latent space. The codebook is integrated into the Transformer decoder as key and value, enabling the model to selectively attend to relevant atomic actions during the decoding process. This design allows the input conditional embedding to be efficiently reconstructed by combining the most appropriate atomic actions based on the cross-attention mechanism in the decoder, leading to more accurate and diverse motion synthesis. Our experiments in Tab. 6 and Fig. 6 show the effectiveness of our codebook design compared to the original CVAE.

Learning Objectives To ensure the effectiveness of the atomic action codebook in generating a wide range of motion sequences, we introduce two objectives for its learning: diversity constraint $\mathcal{L}_{\text{div}}$ and sparsity constraint $\mathcal{L}_{\text{spa}}$. The diversity constraint ensures that the learned atomic actions are diverse and unique, so that the learned atomic actions are diverse enough to represent different actions. By promoting diversity in the codebook, our model can generate more realistic and rich motion sequences that cover a broad spectrum of human actions. Formally, the diversity constraint is formulated as follows:

$$\mathcal{L}_{\text{div}} = \|\boldsymbol{A}\boldsymbol{A}^{\top} - \boldsymbol{I}\|_F, \tag{3}$$

where $\boldsymbol{A}$ is the atomic action codebook, $\boldsymbol{I}$ is the identity matrix, and $\|\cdot\|_F$ is the Frobenius norm. This objective encourages the learned atomic actions to be orthogonal, promoting diversity and preventing codebook redundancy. Our experiments in column "diversity" of Tab. 1, Tab. 2, Tab. 3, and Tab. 4 show that the diversity of our method is higher than the previous approaches and closer to real motions on multiple datasets.

The sparsity constraint promotes the use of a sparse set of atomic actions to represent complex motions, enhancing the atomicity and robustness of the learned atomic actions. This ensures that generated motion sequences are concise and meaningful while maintaining the interpretability and generalizability of the atomic action codebook. We enforce the sparsity constraint by maximizing

Table 1: Quantitative results comparison on the HumanML3D test set. → indicates results are better if they are closer to the real motion, and ± indicates 95% confidence interval.

Method	FID ↓	Diversity →	MultiModality ↑	R Precision (top3) ↑	MultiModal Dist ↓
Real Motion	$0.002^{\pm.000}$	$9.503^{\pm.065}$	-	$0.797^{\pm.002}$	$2.974^{\pm.008}$
Language2Pose [1]	$11.02^{\pm.046}$	$7.676^{\pm.058}$	-	$0.486^{\pm.002}$	$5.296^{\pm.008}$
Text2Gesture [3]	$7.664^{\pm.030}$	$6.409^{\pm.071}$	-	$0.345^{\pm.002}$	$6.030^{\pm.008}$
Hier [10]	$6.532^{\pm.024}$	$8.332^{\pm.042}$	-	$0.552^{\pm.004}$	$5.012^{\pm.018}$
T2M [13]	$\mathbf{0.455^{\pm.003}}$	$9.175^{\pm.002}$	$2.219^{\pm.074}$	$\mathbf{0.736^{\pm.002}}$	$\mathbf{3.347^{\pm.074}}$
MoCoGAN [45]	$94.41^{\pm.021}$	$0.462^{\pm.008}$	$0.019^{\pm.000}$	$0.106^{\pm.001}$	$9.643^{\pm.006}$
Dance2Music [22]	$66.98^{\pm.016}$	$0.725^{\pm.011}$	$0.043^{\pm.001}$	$0.097^{\pm.001}$	$8.116^{\pm.006}$
Ours	$1.691^{\pm.031}$	$\mathbf{9.312^{\pm.011}}$	$\mathbf{2.884^{\pm.130}}$	$0.569^{\pm.004}$	$5.970^{\pm.004}$

the maximal attention values in each cross-attention layer:

$$\mathcal{L}_{\text{spa}} = -\sum_{l}\sum_{h}\max(\boldsymbol{H}_{l,h}), \tag{4}$$

where $\boldsymbol{H}_{l,h}$ is the attention map of the cross-attention for the h-th head in layer l. This loss encourages the model to focus on a few dominant atomic actions when reconstructing motion sequences, leading to a sparser and more interpretable atomic action codebook.

By incorporating the atomic action constraints, the total loss $\mathcal{L}_{\text{total}}$ of our ATOM is as follows:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{CVAE}} + w_{\text{div}}\mathcal{L}_{\text{div}} + w_{\text{spa}}\mathcal{L}_{\text{spa}}, \tag{5}$$

where w_{div} and w_{spa} are weighting hyperparameters.

3.4 Masked Motion Modeling Curriculum Learning

Drawing inspiration from the effective representation learning of masked image autoencoders [16], we introduce masked motion modeling, a technique that involves temporally masking a random portion of the input motion sequence at a ratio r, and subsequently requiring the model to reconstruct the entire motion sequence. Masked motion modeling is essential in the motion synthesis task because it encourages the model to learn robust, context-aware motion representations by forcing it to fill in the missing information. This approach results in a more robust and generalized understanding of the underlying motion structure, ultimately leading to more realistic motion synthesis.

To further enhance the learning process, we incorporate curriculum learning by progressively increasing the mask ratio as the training progresses according to a growth function $g(l) \in \{r, rl/L, r(l/L)^2, re^{l/L-1}\}$, where l is the current training epoch and L is the total number of epochs. As a result, in the beginning, a lower mask ratio allows the model to learn basic motion patterns and capture fundamental structures in the motion data. As the mask ratio increases, the model is exposed to more challenging and complex motion sequences, promoting its ability to infer missing information and makes better use of the learned atomic actions. This curriculum strategy enables a more effective and stable learning experience, ultimately improving motion synthesis performance.

4 EXPERIMENTS

We evaluate our method in three different settings: *text-to-motion*, *action-to-motion*, and *zero-shot action-to-motion*. Text-to-motiona and zero-shot action-to-motion are to generate human motion given an input text prompt; while action-to-motion generates motion given an input action class in the form of one-hot label.

Dataset Five datasets are used for the experiments. For the text-to-motion evaluation, we use the HumanML3D [13] and KIT [36] datasets. The HumanML3D dataset [13], a recent development, is generated through the reannotation of AMASS [31] and HumanAct12 [15] datasets. It consists of 14,616 motions accompanied by 44,970 textual descriptions. The KIT dataset [36] consists of 3,911 motions and the corresponding descriptions.

For the action-to-motion evaluation, we use two different datasets: HumanAct12 [15], UESTC [18], and NTU94 [25, 40]. The UESTC dataset [18] comprises 25K motion sequences spanning 40 classes. HumanAct12 [15] contains 1,191 motions across 12 categories.

We further introduce a novel setting: zero-shot action-to-motion, where the action labels are completely unseen during training. NTU94 [25, 40] is used for this purpose. Specifically, it is a 94-class single-person subset of the NTU RGB+D 120 dataset [25, 40], excluding the 26 categories of multi-person actions. The NTU94 dataset consists of 89K sequences of human motions, and we randomly choose 63 classes as the seen classes for training, and use the remaining 31 classes for the zero-shot action-to-motion synthesis evaluation.

Evaluation Metrics For the text-to-motion evaluation, five metrics are employed, as outlined in [13]: *Fréchet Inception Distance (FID)* quantifies the disparity between generated and ground truth motion distributions in the latent space; *Diversity* evaluates the variation in the generated motion distribution; *Multimodality* measures the average variance given a single text prompt; *R-Precision* and *Multimodal-Dist* assess the relevance of generated motion to the textual prompt. For the R-Precision, we use the top 3 by default. As our primary goal is to improve the motion synthesis quality, we use FID as the prior metric.

For the action-to-motion evaluation, we use four metrics: FID, Diversity, Multimodality and classification accuracy from a pretrained motion classifier.

Implementation Details Our ATOM is trained with the AdamW optimizer for 50K epochs, with an initial learning rate of 10^{-4} decay at the 40K-th step by a factor of 10. All of the hyperparameters are determined via a grid search: $w_{\text{KL}} = w_{\text{div}} = w_{\text{spa}} = 10^{-2}$. We set the hidden dimension $D = 512$, the number of Transformer encoder layers and decoder layers to 8, and the number of attention head $h = 8$. We set the number of atomic actions $N = 256$ for the action-to-motion task, and $N = 1024$ for the text-to-motion task. We follow existing methods [42] to use CLIP [37, 43] for the

Table 2: Quantitative results comparison on the KIT test set.

Method	FID ↓	Diversity →	MultiModality ↑	R Precision (top3) ↑	MultiModal Dist ↓
Real Motion	$0.031^{\pm.004}$	$11.08^{\pm.097}$	-	$0.779^{\pm.006}$	$2.788^{\pm.012}$
Language2Pose [1]	$6.545^{\pm.072}$	$9.073^{\pm.100}$	-	$0.483^{\pm.005}$	$5.147^{\pm.030}$
Text2Gesture [3]	$12.12^{\pm.183}$	$9.334^{\pm.079}$	-	$0.338^{\pm.005}$	$6.964^{\pm.029}$
Hier [10]	$5.203^{\pm.107}$	$9.563^{\pm.072}$	-	$0.531^{\pm.007}$	$4.986^{\pm.027}$
T2M [13]	$2.770^{\pm.109}$	$10.91^{\pm.119}$	$1.482^{\pm.065}$	$\mathbf{0.693^{\pm.007}}$	$\mathbf{3.401^{\pm.008}}$
MoCoGAN [45]	$82.69^{\pm.242}$	$3.092^{\pm.043}$	$0.250^{\pm.009}$	$0.063^{\pm.003}$	$10.47^{\pm.012}$
Dance2Music [22]	$115.4^{\pm.240}$	$0.241^{\pm.004}$	$0.062^{\pm.002}$	$0.086^{\pm.003}$	$10.40^{\pm.016}$
Ours	$\mathbf{0.472^{\pm.029}}$	$\mathbf{10.957^{\pm.092}}$	$\mathbf{2.049^{\pm.086}}$	$0.390^{\pm.006}$	$9.161^{\pm.027}$

Table 3: Quantitative results comparison on the UESTC dataset.

Method	FID (train) ↓	FID (test) ↓	Accuracy ↑	Diversity →	MultiModality →
Real Motion	$2.92^{\pm.26}$	$2.79^{\pm.29}$	$0.988^{\pm.01}$	$33.44^{\pm.320}$	$14.16^{\pm.06}$
Action2Motion [15]	$21.02^{\pm2.51}$	$24.08^{\pm2.17}$	$0.889^{\pm.01}$	$30.47^{\pm.33}$	$13.46^{\pm.03}$
ACTOR [34]	$20.49^{\pm2.31}$	$23.43^{\pm2.20}$	$0.911^{\pm.00}$	$31.96^{\pm.36}$	$\mathbf{14.66^{\pm.03}}$
INR [6]	$9.55^{\pm.06}$	$15.00^{\pm.09}$	$\mathbf{0.941^{\pm.00}}$	$31.59^{\pm.19}$	$14.68^{\pm.07}$
Ours	$\mathbf{6.68^{\pm.04}}$	$\mathbf{9.67^{\pm.17}}$	$0.934^{\pm.01}$	$\mathbf{32.22^{\pm.13}}$	$15.43^{\pm.06}$

Table 4: Quantitative results comparison on the HumanAct12 dataset.

Method	FID (train) ↓	Accuracy ↑	Diversity →	MultiModality →
Real Motion	$0.09^{\pm.01}$	$0.997^{\pm.10}$	$6.85^{\pm.05}$	$2.45^{\pm.04}$
Action2Motion [15]	$2.45^{\pm.08}$	$0.923^{\pm.02}$	$7.03^{\pm.04}$	$2.87^{\pm.04}$
ACTOR [34]	$0.12^{\pm.00}$	$0.955^{\pm.08}$	$\mathbf{6.84^{\pm.03}}$	$2.53^{\pm.02}$
INR [6]	$0.09^{\pm.00}$	$0.973^{\pm.00}$	$6.88^{\pm.05}$	$2.57^{\pm.04}$
Ours	$\mathbf{0.09^{\pm.01}}$	$\mathbf{0.976^{\pm.01}}$	$6.82^{\pm.02}$	$\mathbf{2.52^{\pm.03}}$

Table 5: Quantitative results comparison on the NTU94 dataset. Closed-set (seen classes) and zero-shot (unseen classes) synthesis results are respectively reported.

Method	Seen Classes		Unseen Classes	
	FID (train) ↓	Acc.↑	FID (train) ↓	Acc.↑
Random Generator	$315.55^{\pm0.16}$	$0.016^{\pm.00}$	$319.53^{\pm0.92}$	$0.032^{\pm.00}$
Action2Motion [15]	$138.62^{\pm0.02}$	$0.817^{\pm.00}$	-	-
ACTOR [34]	$136.72^{\pm0.03}$	$0.823^{\pm.00}$	-	-
Action2Motion [15] w/ CLIP	$131.48^{\pm0.01}$	$0.833^{\pm.00}$	$221.57^{\pm0.21}$	$0.108^{\pm.00}$
ACTOR [34] w/ CLIP	$128.80^{\pm0.02}$	$0.835^{\pm.00}$	$205.80^{\pm0.24}$	$0.116^{\pm.00}$
Ours	$\mathbf{42.30^{\pm0.01}}$	$\mathbf{0.869^{\pm.00}}$	$\mathbf{112.95^{\pm0.13}}$	$\mathbf{0.153^{\pm.00}}$

language embedding extraction. For the text-to-motion task, we generate 120 frames, and for the aciton-to-motion task we generate 60 frames. We follow existing methods [13, 15, 34] to select the classifier for metric computation for each task. For the zero-shot action-to-motion task, we use a pretrained ST-GCN classifier [53] to compute the metrics.

4.1 Text-to-motion Evaluation

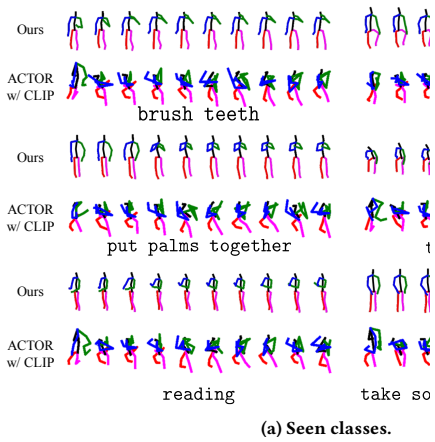
HumanML3D We compare our ATOM with existing methods on the HumanML3D dataset in Tab. 1. Our approach achieves strong FID, signifying a closer match between generated and ground truth motion distributions. Moreover, our ATOM displays superior Diversity and MultiModality scores, showcasing its ability to generate a wide range of motion sequences. Our method is also competitive in terms of R Precision and MultiModal Dist, indicating good alignment between motion and language. Collectively, these results emphasize the strengths of ATOM in generating high-quality and diverse human motion sequences, surpassing previous methods across various aspects.

KIT The results on the KID dataset further emphasize the superiority of our ATOM in motion synthesis. As illustrated in Tab. 2, our method significantly outperforms previous approaches in terms of FID, Diversity, and MultiModality, indicating a substantial reduction in the discrepancy between generated and ground truth motion distributions and the ability to produce diverse motions. Although our ATOM demonstrates promising synthesis outcomes, there is still room for improvement in aligning textual descriptions with the generated motion, which could be addressed by incorporating more advanced language models for language embedding. Overall, these results highlight ATOM's effectiveness in generating high-quality and diverse human motion sequences.

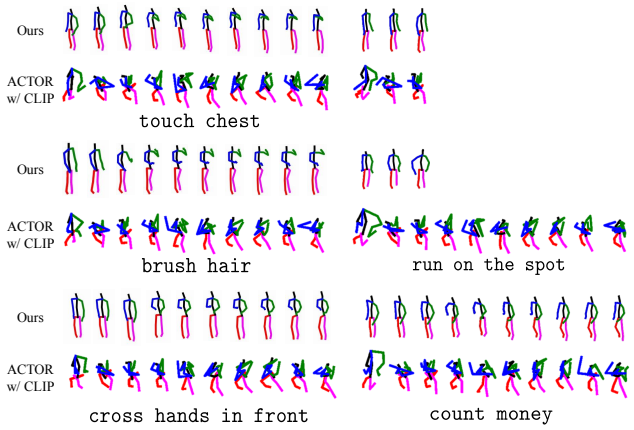
4.2 Action-to-motion Evaluation

UESTC Our experimental results on the UESTC dataset provide a compelling demonstration of ATOM's effectiveness in the action-to-motion synthesis task, as shown in Tab. 3. Among the four metrics, FID is the most important indicator in evaluating the overall synthesis quality. Our ATOM achieves the lowest FID on the training and testing subsets, significantly outperforming the best results from previous methods. This demonstrates that our method generates more realistic motion sequences compared to the competing methods. In terms of accuracy, ATOM attains a strong accuracy of 0.934, indicating that our method is accurate in generating motion sequences that correspond to the given action label. Regarding the diversity of the generated motions, ATOM outperforms previous methods in diversity and achieves competitive multi-modality score. These results illustrate our method's capability to generate realistic, diverse and rich motion sequences.

HumanAct12 The experimental results on the HumanAct12 dataset in Tab. 4 further validate the effectiveness of our ATOM in the action-to-motion synthesis task. Our method and INR [6] achieve the same lowest FID of 0.09, closely approximating the real motion score of 0.09. This consistency between the UESTC and HumanAct12 datasets highlights the robustness of our method in



(a) Seen classes.



(b) Unseen classes.

Figure 4: Zero-shot action-to-motion qualitative results comparison with ACTOR [34] on the NTU94 dataset.

generating realistic motion sequences. Our ATOM also achieves the highest classification accuracy, showing its ability to generate motions that correspond to the action labels across different datasets. We also achieve promising diversity and multi-modality scores. Overall, the consistent and superior performance of ATOM across both UESTC and HumanAct12 datasets emphasizes its effectiveness and robustness in the action-to-motion synthesis task.

4.3 Zero-shot Action-to-motion Synthesis

Apart from the common text-to-motion and action-to-motion settings, we introduce a novel setting: zero-shot action-to-motion synthesis, where the action labels for inference are entirely unseen during training. This new setting aims to evaluate the model's ability to generalize to unseen actions, crucial for practical applications with novel action descriptions. Assessing performance in the zero-shot action-to-motion setting enables a better understanding of the model's potential to synthesize realistic and diverse human motion sequences in real-world scenarios.

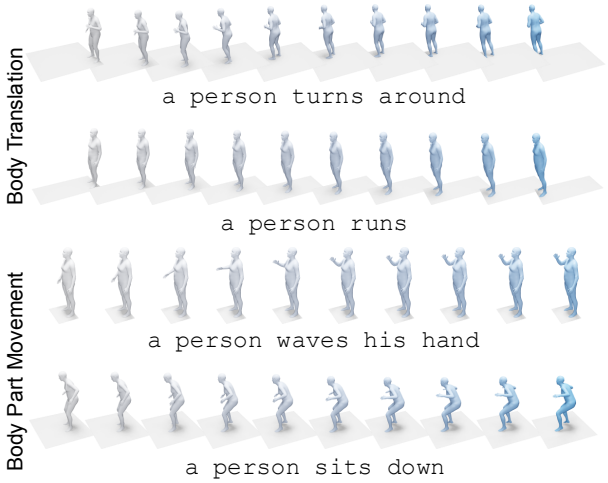


Figure 5: Visualization of learned atomic actions on the HumanML3D dataset. For each textual prompt, we visualize the atomic action that has the highest attention value to it.

We evaluate our zero-shot action-to-motion synthesis performance on the NTU94 dataset, and the results are listed in Table 5. Two previous methods are included for comparison, *i.e.*, ACTOR [34] and Action2Motion [15]. Note they take as input one-hot label vectors to output motion sequences of certain classes, thus they cannot synthesize actions of unseen classes (denoted as "–" in Table 5). To achieve zero-shot synthesis, we replace the one-hot label vector input in with CLIP text embedding, denoted as "w/ CLIP". We make the following observations from the results: (1) Our ATOM significantly outperforms previous methods in both seen class and unseen class synthesis tasks. Surprisingly, our FID on unseen classes is even lower than the ACTOR's FID on seen classes, which demonstrates the effectiveness of our method. (2) Our method exhibits lower variance compared with previous methods, indicating a more stable synthesis results. (3) Though previous methods show promising results on small-scale human motion datasets (*e.g.*, NTU13 and HumanAct12), they show poor capability in large-scale datasets such as NTU94.

We further compare the qualitative results with ACTOR [34] w/ CLIP in Fig. 4. First, our ATOM can successfully generate seen classes, while ACTOR shows poor qualitative results, coinciding previous findings [41] that several previous methods only work well on simple human motion datasets consisting of ~ 10 classes. The high FID and high accuracy of previous methods [15, 34] show that they focus on learning trivial motions for classification, instead of human-recognizable motions. Second, we observe our method generates realistic motions on unseen classes when there are certain body parts in the textual description, *e.g.*, "`touch chest`", "`brush hair`", or "`cross hands in front`". In contrast, ACTOR cannot generate meaningful human motions in such cases.

4.4 Ablation Study

Atomic Action First, we quantitatively analyze the effect of atomic actions in Tab. 6. As the number of atomic actions increases, the FID decreases, and both the Diversity and R Precision improve,

Table 6: Ablation study on the atomic action codebook on the KIT dataset. The variant of model w/o codebook is realized by implementing the decoder with a Transformer encoder, where the query, key and value are the same.

#Atom	$\mathcal{L}_{\text{div}}$	$\mathcal{L}_{\text{spa}}$	FID ↓	Diversity →	R Precision ↑
CVAE baseline			$6.87^{\pm.021}$	$9.102^{\pm.072}$	$0.280^{\pm.024}$
256	-	-	$4.57^{\pm.031}$	$9.204^{\pm.038}$	$0.295^{\pm.039}$
512	-	-	$2.42^{\pm.023}$	$9.673^{\pm.057}$	$0.303^{\pm.034}$
1024	-	-	$1.43^{\pm.038}$	$10.327^{\pm.052}$	$0.337^{\pm.014}$
2048	-	-	$1.47^{\pm.033}$	$10.311^{\pm.062}$	$0.356^{\pm.023}$
1024	✓	-	$0.873^{\pm.021}$	$10.937^{\pm.069}$	$0.381^{\pm.003}$
1024	-	✓	$0.732^{\pm.039}$	$10.688^{\pm.073}$	$0.384^{\pm.012}$
1024	✓	✓	$\mathbf{0.472^{\pm.029}}$	$\mathbf{10.957^{\pm.092}}$	$0.390^{\pm.006}$

Table 7: Ablation study masked motion modeling curriculum learning on the KIT dataset.

Mask ratio r	Learning Scheme	FID ↓	Diversity →	R Precision ↑
w/o masked modeling		$2.48^{\pm.031}$	$10.342^{\pm.031}$	$0.345^{\pm.021}$
25%	$g(l) = r$	$1.22^{\pm.030}$	$10.473^{\pm.037}$	$0.344^{\pm.008}$
50%	$g(l) = r$	$0.76^{\pm.049}$	$10.659^{\pm.051}$	$0.372^{\pm.009}$
75%	$g(l) = r$	$1.48^{\pm.033}$	$10.551^{\pm.034}$	$0.388^{\pm.013}$
50%	$g(l) = rl/L$	$\mathbf{0.472^{\pm.029}}$	$\mathbf{10.957^{\pm.092}}$	$\mathbf{0.390^{\pm.006}}$
50%	$g(l) = r(l/L)^2$	$0.621^{\pm.031}$	$10.683^{\pm.041}$	$0.373^{\pm.009}$
50%	$g(l) = re^{l/L-1}$	$0.583^{\pm.024}$	$10.590^{\pm.057}$	$0.387^{\pm.007}$

indicating better motion generation quality. However, further increasing the number of atomic actions to 2048 does not yield substantial improvements. The combination of diversity and sparsity constraints leads to the best performances, suggesting that they are essential components for generating high-quality, diverse human motion sequences in our model.

We further visually analyze the learned atomic actions in Fig. 5. Upon examination, we find that these atomic actions primarily represent two types of motion: body translation and specific body part movements. Body translation atomic actions typically involve motions like running, or actions involve directional changes, where the entire body is engaged in coordinated movement, see the first two examples of Fig. 5. On the other hand, specific body part movements focus on the motion of a particular body part, such as waving a hand, or sitting down, see the last two examples of Fig. 5. This distinction highlights the versatility of the learned atomic actions, as they capture both global and local motion patterns. As a result, our approach can effectively synthesize complex human motions by combining these diverse atomic actions in a meaningful and contextually appropriate manner.

Masked Motion Modeling Curriculum Learning The results of the masked motion modeling curriculum learning are illustrated in Tab. 7. When compared to the baseline without masked modeling, incorporating masked modeling consistently leads to a reduction in FID and an improvement in Diversity and R Precision, and achieves the best results at a mask ratio of 50%, indicating the benefit of this approach for motion generation quality. Moreover, the learning scheme that follows a linear progression $g(l) = rl/L$ demonstrates the best performance across all metrics, suggesting that a gradual exposure to the complexity of motion reconstruction during training is an effective strategy for enhancing the model's synthesis capabilities.

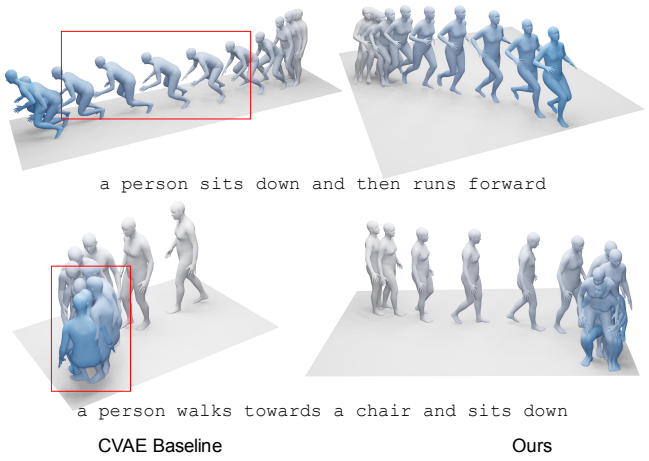


Figure 6: Qualitative comparison between our CVAE baseline and our ATOM.

Qualitative Results We showcase qualitative results in Fig. 6, highlighting the differences between our ATOM and the CVAE baseline. In the first example, the human figure transitioning between "`sitting`" and "`running`" appears tilted in the CVAE result. For the second CVAE example, the person continues to rotate even after sitting down. In contrast, our final ATOM generates a more natural and realistic motion sequences, demonstrating the effectiveness of our proposed atomic action and curriculum learning.

5 CONCLUSION

We present ATOM, a novel approach for language-guided human motion synthesis that addresses the limitations of previous methods by leveraging atomic action decomposition and a curriculum learning strategy. ATOM's transformer-based CVAE framework effectively decomposes complex actions into atomic components, allowing for the generation of diverse and coherent motion sequences. The incorporation of curriculum learning further enhances the model's ability to capture dependencies in motion sequences, resulting in smooth and natural motion transitions between different actions. Our comprehensive evaluation demonstrates ATOM's superior performance over existing approaches in the text-to-motion and action-to-motion synthesis tasks.

ACKNOWLEDGMENTS

This work is supported in part by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR001120C0124, the National Science Foundation (NSF) IIS-2008532, NSF Award #1846076, and the Institute of Education Sciences (IES), U.S. Department of Education (ED) through Award #2229873. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the DARPA, the NFS, the IES, or the ED.

REFERENCES

[1] Chaitanya Ahuja and Louis-Philippe Morency. 2019. Language2pose: Natural language grounded pose forecasting. In *2019 International Conference on 3D Vision (3DV)*. 719–728.

[2] Nikos Athanasiou, Mathis Petrovich, Michael J Black, and Gül Varol. 2022. TEACH: Temporal Action Composition for 3D Humans. *arXiv preprint arXiv:2209.04066* (2022).

[3] Uttaran Bhattacharya, Nicholas Rewkowski, Abhishek Banerjee, Pooja Guhan, Aniket Bera, and Dinesh Manocha. 2021. Text2gestures: A transformer-based network for generating emotive body gestures for virtual agents. In *2021 IEEE virtual reality and 3D user interfaces (VR)*. 1–10.

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

[5] Yujun Cai, Yiwei Wang, Yiheng Zhu, Tat-Jen Cham, Jianfei Cai, Junsong Yuan, Jun Liu, Chuanxia Zheng, Sijie Yan, Henghui Ding, et al. 2021. A unified 3d human motion synthesis model via conditional variational auto-encoder. In *ICCV*. 11645–11655.

[6] Pablo Cervantes, Yusuke Sekikawa, Ikuro Sato, and Koichi Shinoda. 2022. Implicit neural representations for variable length human motion generation. In *ECCV*. 356–372.

[7] Ginger Delmas, Philippe Weinzaepfel, Thomas Lucas, Francesc Moreno-Noguer, and Grégory Rogez. 2022. PoseScript: 3D human poses from natural language. In *ECCV*. 346–362.

[8] Andrea Dittadi, Sebastian Dziadzio, Darren Cosker, Ben Lundell, Thomas J Cashman, and Jamie Shotton. 2021. Full-Body Motion From a Single Head-Mounted Device: Generating SMPL Poses From Partial Observations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 11687–11697.

[9] Yong Du, Wei Wang, and Liang Wang. 2015. Hierarchical recurrent neural network for skeleton based action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1110–1118.

[10] Anindita Ghosh, Noshaba Cheema, Cennet Oguz, Christian Theobalt, and Philipp Slusallek. 2021. Synthesis of compositional animations from textual descriptions. In *ICCV*. 1396–1406.

[11] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. *Advances in neural information processing systems* 27 (2014).

[12] Yusuke Goutsu and Tetsunari Inamura. 2021. Linguistic descriptions of human motion with generative adversarial Seq2Seq learning. 4281–4287.

[13] Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. 2022. Generating diverse and natural 3d human motions from text. In *CVPR*. 5152–5161.

[14] Chuan Guo, Xinxin Zuo, Sen Wang, and Li Cheng. 2022. Tm2t: Stochastic and tokenized modeling for the reciprocal generation of 3d human motions and texts. In *ECCV*. Springer, 580–597.

[15] Chuan Guo, Xinxin Zuo, Sen Wang, Shihao Zou, Qingyao Sun, Annan Deng, Minglun Gong, and Li Cheng. 2020. Action2motion: Conditioned generation of 3d human motions. In *ACM MM*. 2021–2029.

[16] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. 2022. Masked autoencoders are scalable vision learners. In *CVPR*. 16000–16009.

[17] Fangzhou Hong, Mingyuan Zhang, Liang Pan, Zhongang Cai, Lei Yang, and Ziwei Liu. 2022. Avatarclip: Zero-shot text-driven generation and animation of 3d avatars. *arXiv preprint arXiv:2205.08535* (2022).

[18] Yanli Ji, Feixiang Xu, Yang Yang, Fumin Shen, Heng Tao Shen, and Wei-Shi Zheng. 2018. A large-scale RGB-D database for arbitrary-view human action recognition. In *ACM MM*. 1510–1518.

[19] Tero Karras, Samuli Laine, and Timo Aila. 2019. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 4401–4410.

[20] SangBin Kim, Inbum Park, Seongsu Kwon, and JungHyun Han. 2020. Motion Retargetting based on Dilated Convolutions and Skeleton-specific Loss Functions. In *Computer Graphics Forum*, Vol. 39. Wiley Online Library, 497–507.

[21] Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. In *ICLR*.

[22] Hsin-Ying Lee, Xiaodong Yang, Ming-Yu Liu, Ting-Chun Wang, Yu-Ding Lu, Ming-Hsuan Yang, and Jan Kautz. 2019. Dancing to music. *NeurIPS* 32 (2019).

[23] Bowen Li, Xiaojuan Qi, Thomas Lukasiewicz, and Philip HS Torr. 2020. Manigan: Text-guided image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7880–7889.

[24] Gen Li, Nan Duan, Yuejian Fang, Ming Gong, and Daxin Jiang. 2020. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 11336–11344.

[25] Jun Liu, Amir Shahroudy, Mauricio Perez, Gang Wang, Ling-Yu Duan, and Alex C Kot. 2019. Ntu rgb+d 120: A large-scale benchmark for 3d human activity understanding. *IEEE TPAMI* (2019), 2684–2701.

[26] Yahui Liu, Marco De Nadai, Deng Cai, Huayang Li, Xavier Alameda-Pineda, Nicu Sebe, and Bruno Lepri. 2020. Describe what to change: A text-guided unsupervised image-to-image translation approach. In *Proceedings of the 28th ACM International Conference on Multimedia*. 1357–1365.

[27] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. 2015. SMPL: A skinned multi-person linear model. *ACM TOG* 34, 6 (2015), 1–16.

[28] Tianyu Luan, Yali Wang, Junhao Zhang, Zhe Wang, Zhipeng Zhou, and Yu Qiao. 2021. Pc-hmr: Pose calibration for 3d human mesh recovery from 2d images/videos. In *AAAI*. 2269–2276.

[29] Tianyu Luan, Yuanhao Zhai, Jingjing Meng, Zhong Li, Zhang Chen, Yi Xu, and Junsong Yuan. 2023. High Fidelity 3D Hand Shape Reconstruction via Scalable Graph Frequency Decomposition. In *CVPR*. 16795–16804.

[30] Thomas Lucas, Fabien Baradel, Philippe Weinzaepfel, and Grégory Rogez. 2022. PoseGPT: quantization-based 3D human motion generation and forecasting. In *ECCV*. 417–435.

[31] Naureen Mahmood, Nima Ghorbani, Nikolaus F Troje, Gerard Pons-Moll, and Michael J Black. 2019. AMASS: Archive of motion capture as surface shapes. In *ICCV*. 5442–5451.

[32] Seonghyeon Nam, Yunji Kim, and Seon Joo Kim. 2018. Text-adaptive generative adversarial networks: manipulating images with natural language. *Advances in neural information processing systems* 31.

[33] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. 2021. Styleclip: Text-driven manipulation of stylegan imagery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2085–2094.

[34] Mathis Petrovich, Michael J Black, and Gül Varol. 2021. Action-conditioned 3d human motion synthesis with transformer vae. In *ICCV*. 10985–10995.

[35] Mathis Petrovich, Michael J Black, and Gül Varol. 2022. TEMOS: Generating diverse human motions from textual descriptions. *arXiv preprint arXiv:2204.14109* (2022).

[36] Matthias Plappert, Christian Mandery, and Tamim Asfour. 2016. The KIT motion-language dataset. *Big data* (2016), 236–252.

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

[38] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-shot text-to-image generation. In *International Conference on Machine Learning*. PMLR, 8821–8831.

[39] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. 2016. Generative adversarial text to image synthesis. In *International conference on machine learning*. PMLR, 1060–1069.

[40] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. 2016. Ntu rgb+d: A large scale dataset for 3d human activity analysis. In *CVPR*. 1010–1019.

[41] Ziyang Song, Dongliang Wang, Nan Jiang, Zhicheng Fang, Chenjing Ding, Weihao Gan, and Wei Wu. 2022. ActFormer: A GAN Transformer Framework towards General Action-Conditioned 3D Human Motion Generation. *arXiv preprint arXiv:2203.07706* (2022).

[42] Guy Tevet, Brian Gordon, Amir Hertz, Amit H Bermano, and Daniel Cohen-Or. 2022. Motionclip: Exposing human motion generation to clip space. In *ECCV*. 358–374.

[43] Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H Bermano. 2022. Human motion diffusion model. *arXiv preprint arXiv:2209.14916* (2022).

[44] Joëlle Tilmanne and Thierry Dutoit. 2010. Expressive gait synthesis using PCA and Gaussian modeling. In *International Conference on Motion in Games*. Springer, 363–374.

[45] Sergey Tulyakov, Ming-Yu Liu, Xiaodong Yang, and Jan Kautz. 2018. Mocogan: Decomposing motion and content for video generation. In *CVPR*. 1526–1535.

[46] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *NeurIPS* 30 (2017).

[47] Jiashun Wang, Huazhe Xu, Jingwei Xu, Sifei Liu, and Xiaolong Wang. 2021. Synthesizing long-term 3d human motion and interaction in 3d scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9401–9411.

[48] Zan Wang, Yixin Chen, Tengyu Liu, Yixin Zhu, Wei Liang, and Siyuan Huang. 2022. Humanise: Language-conditioned human motion generation in 3d scenes. *arXiv preprint arXiv:2210.09729* (2022).

[49] Tianyi Wei, Dongdong Chen, Wenbo Zhou, Jing Liao, Zhentao Tan, Lu Yuan, Weiming Zhang, and Nenghai Yu. 2021. Hairclip: Design your hair by text and reference image. *arXiv preprint arXiv:2112.05142* (2021).

[50] Weihao Xia, Yujiu Yang, Jing-Hao Xue, and Baoyuan Wu. 2021. Tedigan: Text-guided diverse face image generation and manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2256–2265.

[51] Kevin Xie, Tingwu Wang, Umar Iqbal, Yunrong Guo, Sanja Fidler, and Florian Shkurti. 2021. Physics-based Human Motion Estimation and Synthesis from Videos. In *Proceedings of the IEEE/CVF International Conference on Computer*

Vision. 11532–11541.

[52] Sijie Yan, Zhizhong Li, Yuanjun Xiong, Huahan Yan, and Dahua Lin. 2019. Convolutional sequence generation for skeleton-based action synthesis. In *ICCV*. 4394–4402.

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

[54] Ping Yu, Yang Zhao, Chunyuan Li, Junsong Yuan, and Changyou Chen. 2020. Structure-aware human-action generation. In *European Conference on Computer Vision*. Springer, 18–34.

[55] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. 2017. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*. 5907–5915.

[56] Junhao Zhang, Yali Wang, Zhipeng Zhou, Tianyu Luan, Zhe Wang, and Yu Qiao. 2021. Learning dynamical human-joint affinity for 3d pose estimation in videos. *IEEE TIP* (2021), 7914–7925.

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

[58] Yan Zhang, Michael J Black, and Siyu Tang. 2020. Perpetual motion: Generating unbounded human motion. *arXiv preprint arXiv:2007.13886* (2020).