Vision, Modeling, and Visualization (2024)
L. Linsen and J. Thies (Eds.)

COMAND: Controllable Action-aware Manifold for 3D Motion Synthesis

I. Habibie[1], M. Elgharib[1], D. Luvizon[1], B. Thambiraja[2,3], S. Nyatsanga[4], J. Thies[2,3], M. Neff[4] and C. Theobalt[4]

[1] MPI Informatics, Germany, [2]TU Darmstadt, Germany
[3]MPI Intelligent Systems, Germany, [4]UC Davis, USA

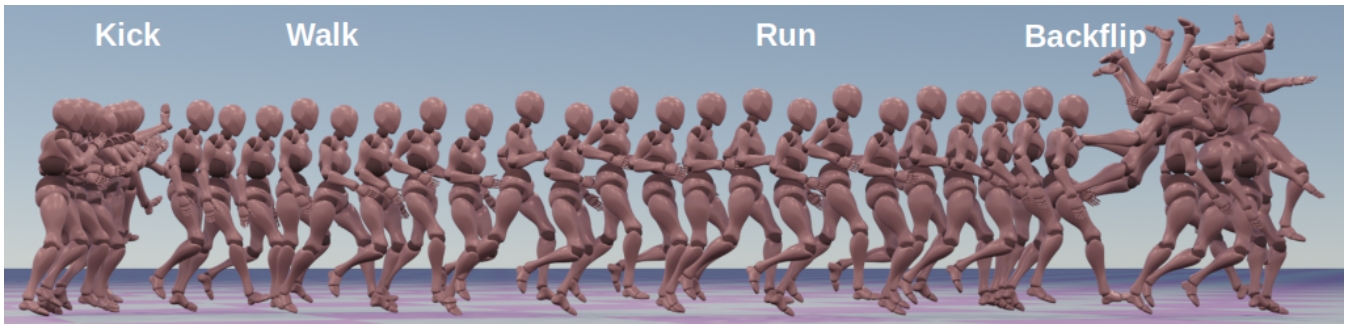


Figure 1: *COMAND is a novel framework for synthesizing 3D motion sequences containing multiple and complex actions. It is based on a novel frequency-based motion manifold which can be constructed without an action-labeled dataset.*

Abstract

We present COMAND, a novel method for controllable multi-action 3D motion synthesis without requiring action-labeled data. Our method can generate a lifelike motion sequence containing consecutive non-locomotive actions such as kicking, jumping, or squatting, without the need for manual blending, enabling an intuitive way to control 3D human animation based on the desired motion types at specified time windows. At the core of our method is a motion manifold based on a periodic parameterization of a motion latent space that allows for unsupervised action clustering of 3D motion, thus allowing action-to-motion synthesis without the need to explicitly train the model on action-labeled datasets. This learned motion manifold has semantic and periodic properties that benefit 3D motion synthesis from action labels and from free-form text input, resulting in a state-of-the-art multi-modal and multi-action 3D motion generation framework. Our study shows that more than 83% and 96% of the users respectively rated COMAND as more natural and better matching the target action sequence when compared to existing methods.

CCS Concepts
• ***Computing methodologies*** → ***Motion capture;***

1. Introduction

Synthesizing realistic 3D motion of humans is an essential task in many computer vision and graphics-related applications. At the same time, synthesizing 3D motion is also known to be a challenging task due to the highly articulated nature of the human body, traditionally requiring many hours of manual work from skilled artists.

Recent developments in data-driven methods have led to significant progress in this area. For example, a considerable number of works have shown the potential of learning-based approaches to synthesize natural-looking 3D motion from directional control [LWB*10, HSK16, HKS17, ZSKS18, HKPP20]. However, many recent 3D motion synthesis approaches focus on locomotion data consisting of a limited number of actions beyond walk-related movements. They often struggle to produce aperiodic actions, such as kicking or cartwheeling, especially when these annotated motion sequences are scarce.

DOI: 10.2312/vmv.20241209

Meanwhile, tremendous progress in generative modeling approaches, such as diffusion models [SDWMG15, HJA20, DN21], has led to a new class of methods capable of producing high-quality synthesis in several domains, including image [RPG*21, RDN*22, SCS*22, RBL*22], audio [KPH*21], 3D mesh [PJBM22, LGT*23, LWH*23], and video [SPH*22, HCS*22]. Naturally, several works have investigated the extension of such approaches to synthesize 3D human animation [AM19, GCO*21, PBV22, ZCP*22]. These models have been shown capable of producing animations that often match the description of the action. However, while synthesized motion sequences are generally plausible, such methods often suffer from catastrophic neglect [CAV*23a], which prevents them from reliably generating 3D motion sequences that consist of more than two consecutive actions. Current text-based approaches are not designed to allow fine-grained control over individual actions, which is a major limitation for many applications.

To address these issues, we propose a new framework for synthesizing 3D human motion from a high-level description of motion labels, e.g., motion action labels or free-form text input. Unlike prior action-based motion synthesis approaches [GZW*20, PBV21], our method can seamlessly generate multiple actions consecutively without the need to manually blend the individual actions. Our method does not require the data to be pre-labeled thanks to our latent space that allows for unsupervised clustering of different types of motion based on their semantically meaningful action. Unlike previous free-form text-based synthesis approaches [TRG*22, DMGT22], in our method the users can specify a certain action at regular time intervals, allowing finer-level control over the timing and duration of each action. The key to our approach is a frequency-based latent space that produces a unique range of values for one particular action of 3D motion for a given time window. In this space, each latent vector representing a motion can be grouped together with other latent vectors of similar actions using off-the-shelf clustering algorithms. The inherent properties of our motion representation allow us to use our frequency-based latent space as an intermediate embedding for the task of text-based 3D motion synthesis. We show that the motion synthesis from text is further improved by retrieving motion features in our frequency-based latent space from the training corpus.

In summary, we propose a new approach that enables interactive and intuitive 3D synthesis with the following contributions:

- A new way to design a latent representation of 3D human motion that is interpretable and can be used to cluster unlabeled 3D motion data into distinct actions, thus aiding action-based synthesis.
- A novel framework that enables interactive and controllable multi-action 3D motion synthesis that can be used even when the label is not initially available.
- A novel action-based motion synthesis model that can be controlled using free-form text inputs or by querying specific action types from a database.

2. Related Work

In the following, we review the related work in learning human motion manifolds and motion synthesis.

2.1. Learning Human Motion Manifold

Neural network-based models can efficiently encode abstract representations of human motion in a scalable manner, making them highly amenable for data-driven workflows in character animation. While prior data-driven approaches [CH05, MK05, GMHP04] can also be used to learn motion manifolds, they tend to require intensive data preprocessing and do not scale for large datasets.

Holden *et al.* [HSKJ15] trained a convolutional autoencoder to encode the entire CMU motion capture dataset [CMU03]. Holden *et al.* [HSK16] combined a latent representation with a feedforward network to achieve 3D motion synthesis and control.

A challenge in 3D motion synthesis is to reduce artifacts such as foot skating. With this goal, Holden *et al.* [HKS17] proposed Phase-Functioned Neural Network (PFNN), which separates the phase information of the locomotion data based on the foot-ground contacts. This phase variable is then used by the neural network to blend its weights accordingly to ensure a high-quality transition during synthesis. DeepPhase [SMK22] extends this idea by learning a phase manifold space over the joints of the whole body, leading to better temporal and spatial alignment when compared to the foot contact-based phase information. If the manifold is trained from a skill-specific dataset, the phase information can be inferred during synthesis to guide the prediction of the 3D pose output for the next time steps. Unfortunately, this assumption generally no longer holds for a large dataset containing various aperiodic actions. In our approach, we propose to decompose the latent information into its phasic and non-phasic components, revealing an interesting behavior that allows us to better cluster the actions within a motion dataset by disregarding the periodic information.

2.2. Action and Text Conditioned Motion Synthesis

Recently, several works have proposed that connect 3D motion with text inputs using a shared embedding [AM19, GCO*21, ZZC*23]. Guo *et al.* [GZZ*22] proposed a temporal variational autoencoder (VAE) combined with a motion length sampler to generate diverse motion samples from text input. Guo *et al.* [GZWC22] leveraged an attention-based GRU network to model the motion tokens enabling text-to-motion prediction in an autoregressive manner. Recently, several works explored using conditional diffusion models for the task of text-to-motion synthesis. Zhang *et al.* [ZCP*22] and Tevet *et al.* [TRG*22] concurrently proposed a Transformer-based diffusion model to perform text-to-motion synthesis, and MoFusion [DMGT22] proposed a CNN-based diffusion approach that can be either conditioned by text and music inputs.

Motion synthesis can be conditioned by categorical labels. Guo *et al.* [GZW*20] proposed a conditional temporal VAE that generates diverse human motion given an action label. Petrovich *et al.* [PBV21] improved on this by proposing a Transformer-based encoder-decoder, optimized with the VAE objective, to generate diverse and variable length motion sequences given an action label. Unlike Guo *et al.* [GZW*20], whose approach is autoregressive with a frame-level latent representation, Petrovich *et al.* [PBV21] leveraged positional encoding from the Transformer architecture and learned sequence-level latent vectors, enabling the model to generate variable-length sequences all at once. However, these two

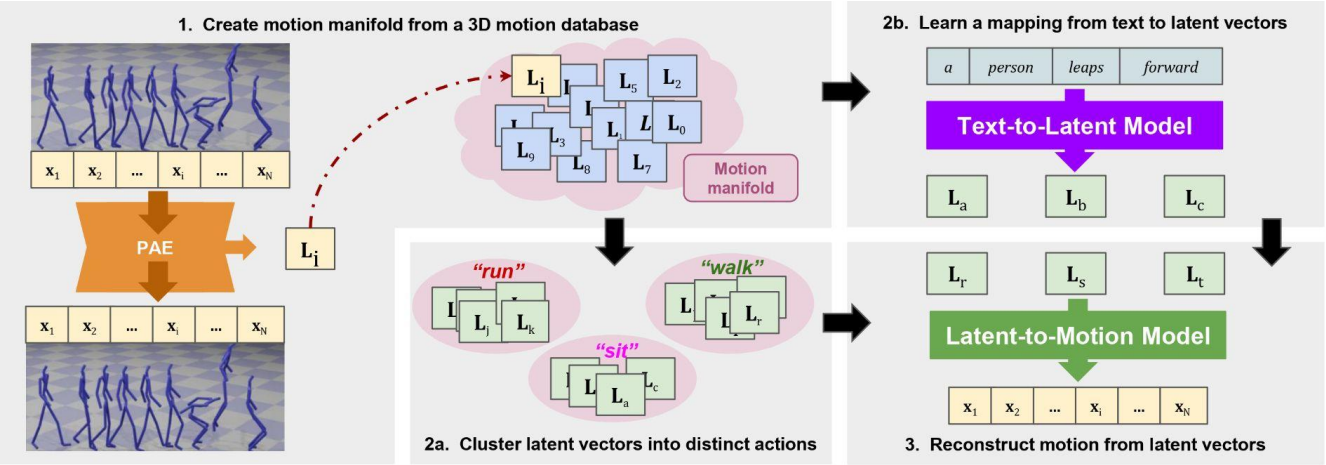


Figure 2: *The pipeline for the proposed multi action-to-motion synthesis approach. In Stage 1, given a collection of 3D motion in a database, we extract the latent vector representations of each 3D motion sequence using a Phase Autoencoder (PAE) network (Sec 3.1). In Stage 2, the user is given two choices: 2a) group the latent vector database into N number of clusters that represent unique actions (Sec 3.2), or 2b) use the latent database to train a Text-to-Latent regressor model in a supervised manner (Sec 3.3.2). Finally, in Stage 3, the latent vector sequence generated from either latent cluster sampling (Stage 2a) or text-to-latent mapping (Stage 2b) is passed into a Latent-to-Motion regressor model to synthesize the desired 3D motion sequence (Sec 3.3.1).*

approaches only condition motion generation on categorical action labels which have limited semantics. Interestingly, Tevet *et al.* [TGH*22] enhanced the limited semantics of short text phrases with CLIP [RKH*21]. Their key insight is to align the embedding space with that of CLIP by using text phrases and rendered images of the motion. This enriched the motion manifold with semantic information from CLIP-space, enabling semantically similar motions to be grouped together in the latent space.

Guiding motion generation with full natural language sentences would be desirable because they have richer semantics and are more descriptive. Guo *et al.* [GZZ*22] proposed a two-stage approach for generating diverse, variable-length motion sequences, given a text description of a string of actions. Petrovich *et al.* [PBV22] improved on this probabilistic approach by leveraging their earlier work that used action labels [PBV21] using a novel Transformer model. In this work, we propose a new way of controlling the synthesis of 3D human motion where the input label can either be directed from a list of free-form text-based action labels or discovered in an unsupervised manner.

3. Approach

Our main goal is to achieve multi-action 3D motion synthesis. To this end, we propose using a neural latent representation of 3D motion that accommodates unsupervised clustering of distinct actions over the database. This representation is then used to achieve action-to-motion synthesis, either from a cluster sample or text input. The pipeline of our proposed approach is depicted in Fig. 2.

3.1. Action-aware Latent Space for 3D Motion

Given a collection of 3D human motion sequences represented by $\mathbf{X}_i = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_i, ..., \mathbf{x}_{T-1}, \mathbf{x}_T\}$, where $\mathbf{x}_i$ is the 3D human pose at frame i, we are interested in extracting a semantically meaningful latent space where unique action types can be automatically separated. While there have been a number of approaches proposed to construct a motion manifold, none of them focuses on designing a representation that can be used to identify distinct actions from a collection of unlabeled data. Naively clustering a latent space obtained by a convolutional neural network [HKS17] or a VAE [KW14] may not lead to an optimal embedding, as their representation is not specifically designed to distinguish the action between motion sequences.

The recently proposed Periodic Autoencoder (PAE) [SMK22] presents a new way of encoding 3D motion data using a parameterized latent neural representation in the frequency space. Given a convolutional encoder $\mathcal{E}(\mathbf{X})$ and a differentiable Fast Fourier Transform (FFT) layer $\mathcal{F}(\cdot)$, each channel of the latent space $\mathbf{H} = \mathcal{F}(\mathcal{E}(\mathbf{X})) \in \mathbb{R}^C$ is parameterized using a sinusoidal function by fitting each channel of the encoder output according to

$$\mathbf{L} = f(T; \mathbf{A}, \mathbf{F}, \mathbf{P}, \mathbf{B}) = \mathbf{A} \cdot \sin(2\pi \cdot (\mathbf{F} \cdot T - \mathbf{P})) + \mathbf{B}, \quad (1)$$

where $\mathbf{H} \approx \mathbf{L}$, $\mathbf{A} \in \mathbb{R}^C$, $\mathbf{F} \in \mathbb{R}^C$, $\mathbf{P} \in \mathbb{R}^C$, and $\mathbf{B} \in \mathbb{R}^C$ refer to the amplitude, frequency, phase, and offset (zero-frequency) parameters of the sinusoid of a given time window T. Given a decoder $\mathcal{D}(\cdot)$, the model is trained in an autoencoding manner using

$$\mathcal{L}_{MSE} = (\mathbf{X} - \mathcal{D}(\mathbf{L}))^2. \quad (2)$$

One interesting property of this representation is that the non-phasic amplitude $\mathbf{A}$, frequency $\mathbf{F}$, and offset $\mathbf{B}$ parameters that are usually near-constant over a short motion window. Their value tends to be consistent for similar actions, making it a good candidate for an action-aware latent space once we discard the phase information $\mathbf{P}$ from the latent space.

While the standard PAE network can be used to extract periodic

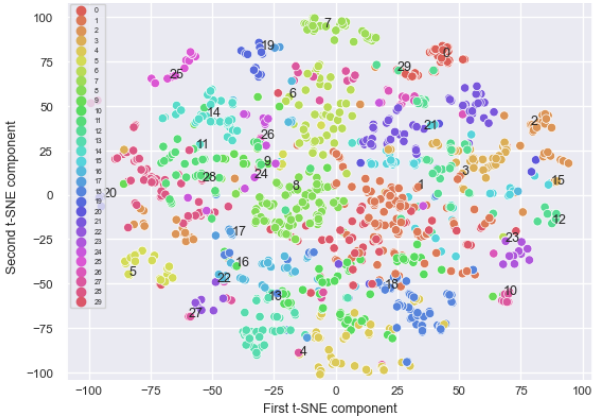


Figure 3: *T-SNE projection of the non-phasic component of our PAE latent space. We apply k-means clustering on a subset of the HumanML3D data [GZZ*22] to obtain the labels (see Tab. 1). The t-SNE plot shows that similar motion tend to be close together in latent space which is beneficial for motion synthesis.*

features on non-periodic motion data, the periodicity characteristic of the motion will no longer be prominent when we apply the autoencoder to the non-repetitive motion. As a result, phase-based latent features will perform poorly when used on large-scale unstructured motion data consisting of numerous types of aperiodic actions. On the other hand, we observe that the value of the non-phasic parameters of a given short motion is usually near-constant for any given type of action. We hypothesize that the non-phasic information acts as a low-dimensional feature that extracts unique motion characteristics of a given action label. Consequently, if the non-phasic value of a certain action can be uniquely identified, then we can make use of such parameterizations as a target latent space that is easy to cluster into distinct actions.

To construct our action-aware latent space, we train the PAE network on short 3D motion sequences of the same length, e.g. 2 seconds. While each latent vector captures the information of fixed-length motion, as we later discuss in Section 3.3, a variable length motion can be reconstructed back into the 3D motion space by passing a sequence of latent vectors into a Latent-to-Motion network.

3.2. Clustering the Action-aware Latent Space

To examine our hypothesis about the uniqueness of the non-phasic components of the periodic latent space, we apply k-means clustering in this latent space using a subset of the HumanML3D [GZZ*22] motion dataset that contains various actions. This subset is chosen to contain motion sequences that are around 2-4 seconds long. In Fig. 3, we visualize the clustering result using t-SNE [VdMH08], which projects the high-dimensional non-phasic latent vectors consisting of the amplitude **A**, frequency **F**, and offset **B** parameters into a 2D space. Action labels are shown in Tab. 1. The t-SNE plot of the motion reveals that data points that belong to the same action tend to occupy the same region in the latent space. Note that this structure is achieved without explicitly optimizing the model to have an action-aware latent space.

(a) *Running (13)* **(b)** *Jumping jack (23)* **(c)** *Raise left hand (0)*

Figure 4: *Motion snippets generated by sampling three different clusters from our PAE latent space.*

Table 1: *The action label annotation of the PAE latent clusters was obtained using the K-Means algorithm, with K=30. The outliers were manually removed. *Mixed label indicates clusters where the dominant action type cannot be clearly identified.*

ID	Action label	ID	Action label
0	waving/moving left hand	15	grabbing/punching
1	jump	16	kick
2	raise arm	17	mixed*
3	left hand	18	walking variations
4	walks forward	19	sitting and steering
5	sitting down	20	punching/throwing
6	jump	21	applause/arms crossing
7	sitting while moving hands	22	cartwheel/backflip/roundhouse kick
8	waving/moving right hand	23	jumping jack
9	raise hand	24	lifting arm
10	running/long stride walk	25	stretch arms
11	lifting objects with hand	26	raise/cross arms
12	standing and raising hand	27	leap
13	running/jogging	28	stretch/raise/move hands
14	mixed*	29	raising/standing up

To validate the uniqueness of the action within every cluster, we visually inspect the clusters by reconstructing the latent vectors back into the 3D motion space. We then randomly sample the elements and manually inspect the results. Fig. 4 shows sampled frames from different action clusters.

3.3. Action-to-Motion Framework

3.3.1. Latent-to-Motion Regression

So far, we have discussed how the periodic latent space is constructed to cluster different types of actions. However, the PAE decoder is not designed to reconstruct motion sequences of varying lengths. Since our goal is to reconstruct 3D motion from any sampled latent vectors from the cluster, we ideally want the decoder to generate different types of actions within a single pass without the need for manually blending different motion sequences. To address this issue, we propose a separate Latent-to-Motion network that maps a variable-length sequence of latent vectors into 3D motion. We leverage a dual Encoder Transformer-based model [PBV22] as the base architecture for the Latent-to-Motion network.

Since by design the PAE latent vector encodes motion information from a time window of 2 seconds, there exist scenarios where a particular action may not have been completed before the next ac-

tion from the following vector begins. For example, if an animator intends to generate a kick sequence followed by a walk action, the kick may not be complete if there is not enough time between the two actions. To improve its usability for semi-automated synthesis tasks, the proposed Latent-to-Motion model (see Fig. 5) is also designed to take a blank vector containing zero values as input. Fig.1 shows a synthesis of a motion generated by the regressor.

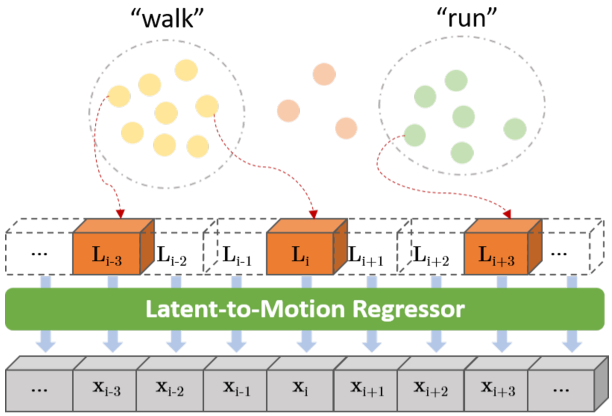


Figure 5: *Our Latent-to-Motion regressor maps a sequence of latent vectors* $\{\mathbf{L}_1, \mathbf{L}_2, ..., \mathbf{L}_J\}$ *into a 3D motion* $\{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_T\}$. *Each latent vector can be sampled from a certain cluster or generated through other means (see Sec. 3.3.2). Every two consecutive latent vectors are 1 second apart in the motion space.*

3.3.2. Text-to-Latent Regression

So far, our method has shown to be capable of performing action-controllable 3D motion synthesis by forwarding the appropriate sample from a given action cluster to the Latent-to-Motion regressor. At the same time, it is often desirable if the actions can be generated from a free-form text input. However, synthesizing 3D motion from a free-form text description is a challenging task as there can be multiple correct solutions that satisfy the input. We propose to resolve this issue by predicting the latent space from text input using another Transformer network. To achieve the final motion output, we pass the output of the Text-to-Latent network as input for the Latent-to-Motion model.

3.4. Text-to-Motion Architecture

Both the proposed Latent-to-Motion and Text-to-Latent models use a similar Transformer-based architecture to regress their target predictions, but they also have some subtle differences. For the Text-to-Latent regressor, we use the dual encoder Transformer model [PBV21, PBV22] originally designed for action-to-motion and text-to-motion synthesis tasks, and we modify it to receive the latent vector sequence as input. In contrast to the original Transformer implementation used for language translation tasks, this particular implementation is not autoregressive and predicts the whole output sequence within a single inference pass. The latent encoder of the Text-to-Latent regressor consists of a combination of self-attention, layer norm, pointwise feed-forward, and dropout layers commonly found in a vanilla Transformer architecture. For the Text encoder of the Text-to-Latent regressor, we used a pre-trained DistilBERT [SDCW19] model to encode the text input into a sequence of embedding vectors before passing it into the standard Transformer encoder model. We use 6 attention blocks with 4 attention heads for both encoders. We only keep the first element in the output embedding sequence of each encoder. We use the reconstruction errors from both the Text-to-Latent and the Latent-to-Latent paths as the main objective. In addition, we also apply an embedding similarity loss between the latent vectors produced by the last layer of the encoders with a weight factor of $1e-5$. We used smooth L1 loss in every loss term.

3.4.1. Matching-based Text-to-Motion Synthesis

Our proposed text-to-motion design can be seen as a soft quantization where the model first predicts an intermediate representation of the motion. However, despite our target non-phasic latent space being a more robust prediction target than the original 3D motion space, it is still possible that the model may fail to properly map the input into the target domain due to regression error. At the same time, all 3D motion sequences in the training data can be converted into a collection of latent vectors. In contrast to the latent vectors predicted by the Text-to-Latent regression, which may introduce some error, these vectors lie in the actual manifold of the training data. Since such vectors can be seen as the *ground truth* for the latent representation, we can use this collection as a database of latent vectors to "project" our predicted vectors against. We do this by replacing the prediction with the closest vector in the collection using the nearest neighbor algorithm. This approach can be seen as a post-training hard quantization of the latent space since each predicted vector is "quantized" into one of the training vectors.

4. Experiment and Evaluation

In this section we perform evaluations considering two scenarios. First, we evaluate our method on the task of multi-action motion synthesis. Second, we compare our approach with existing free-form text-based methods. Even though this setup is not the main goal of our paper, most recent methods are evaluated on text-based benchmarks, therefore, it provides a good comparison with existing approaches. Finally, we show the superiority of our method in qualitative results.

4.1. Multi-Action Motion Synthesis

We assess the capability of our model to generate 3D human motion that contains multiple action types within a single sequence. Here, we compare our approach against TEACH [APBV22], the state-of-the-art method for multi-action 3D motion synthesis. In addition, we also compare the performance of our method against MDM [TRG*22], one of the state-of-the-art diffusion-based free-form text-based motion synthesis approaches. One way to evaluate the performance would be to compare each method on the subset of an action-to-motion dataset that contains multi-action sequences. However, we find the available amount of action transitions in such multi-action datasets, e.g. BABEL [PCA*21], to be insufficient to properly assess the ability of multi-action approaches to generalize to various action combinations.

Furthermore, while our method and MDM were trained on HumanML3D [GZZ*22], TEACH was trained on the BABEL dataset [PCA*21], making it difficult to perform a fair evaluation on the existing annotated datasets. To address this issue, we propose a qualitative evaluation strategy where we compare the synthesis from various possible permutations of the input text prompt from a set of pre-selected actions found in the existing datasets. Three different sets of text prompts were thus prepared to accommodate the input styles of each model. Please refer to our supplementary material for more details on the text prompts. In total, we generate combinations of multi-action sequences based on 20 distinct and commonly found activities in the HumanML3D or BABEL dataset.

Table 2: *User study comparing our action-based synthesis quality against TEACH and MDM. Most users generally agree that our approach is preferable compared to the competing methods, both in terms of naturalness and its correctness to synthesize the specified sequence of action inputs.*

Method	Naturalness	Action Matching
Ours vs. TEACH	96.94%	96.53%
Ours vs. MDM	83.88%	96.12%

We evaluate the performance of the models based on the quality of their output in generating sequences that contain two or three consecutive actions. For ours and TEACH, we randomly combine 2 or 3 sequences lasting 3 seconds each, which results in 6 or 9 seconds of motion. Since MDM is trained with free-form text data without duration conditioning, its output does not have a specific duration time per action. Nonetheless, since the model is trained on sequences that are at most 9.8 seconds long, MDM is able to show consecutive two- or three-action sequences within the given time window of 6 or 9 seconds. For MDM, we combine the consecutive actions into a single sentence. Such consecutive actions are frequent in the HumanML3D dataset.

We performed paired comparison tests to evaluate the synthesis quality of our method against the competing approaches of TEACH and MDM. For each of the ten randomly sampled action prompts, we produced stimuli pairs for each combination of TEACH or MDM against our model (i.e. TEACH-Ours, MDM-Ours). This results in a total of 20 comparisons. The clip order was randomized. For each pair, participants were required to select one preferred clip for each of two criteria (Two-alternative forced choice (2AFC) design): 1) the naturalness of the synthesized motion, and 2) the match between the given action description and the synthesized motion. The user study took around 5 minutes to complete. It was run online using a convenience sample of 49 participants. Most participants agree that our model produces a more natural and higher matching with respect to the input action. Our proposed approach is rated to be better than any competing approaches by at least 80% on any comparison scenarios. The results are shown in Tab. 2. According to Exact Binomial Tests, the results are highly significant in all cases ($p < .0001$).

4.2. Text-to-Motion Synthesis

We also compare our text-to-motion approach against prior works on the HumanML3D dataset [GZZ*22]. The dataset provides two main evaluation metrics for the text-to-motion synthesis task. The first metric is R-precision, where each of the generated 3D motions is compared and ranked against a collection of text descriptions containing its ground truth label as well as 31 randomly sampled labels. This comparison is performed in the latent space of a pre-trained model that measures the similarity between the given motion and text input based on their Euclidean distance. The retrieval is considered successful if the ground truth falls within the top-K labels. The second metric is the Frechet Inception Distance (FID) adopted from Guo *et al.* [GZW*20], which compares the distribution of the synthesized 3D motion against the reference ground truth data. Additionally, the variance of the generation is also assessed through the diversity metric.

Table 3: *Quantitative comparison of the text-to-motion task on the HumanML3D dataset. (*) denotes methods that use Inverse Kinematics (IK) to obtain the 6D joint angle representation in the over-parameterized HumanML3D evaluation set required for evaluation. The symbols ↑, ↓, and → indicate that a score is considered better if it is higher, lower, or closer to the ground truth compared to other methods, respectively.*

Method	R Prec. ↑ (top 3)	FID↓	Div. →
Real	0.797	0.002	9.503
JL2P [AM19]	0.486	11.020	7.676
Text2Gest. [BRB*21]	0.345	7.664	6.409
MotionDiffuse [ZCP*22]	0.782	0.630	9.410
TM2T [GZWC22]	0.729	1.501	8.589
T2M-GPT [ZZC*23]	0.775	0.116	9.761
MLD [CJL*22]	0.772	0.473	9.724
MDM [TRG*22]	0.611	0.544	9.559
MoFusion* [DMGT22]	0.492	–	8.820
Ours*	0.710	1.156	9.277
Ours* (after matching)	0.662	0.875	9.562

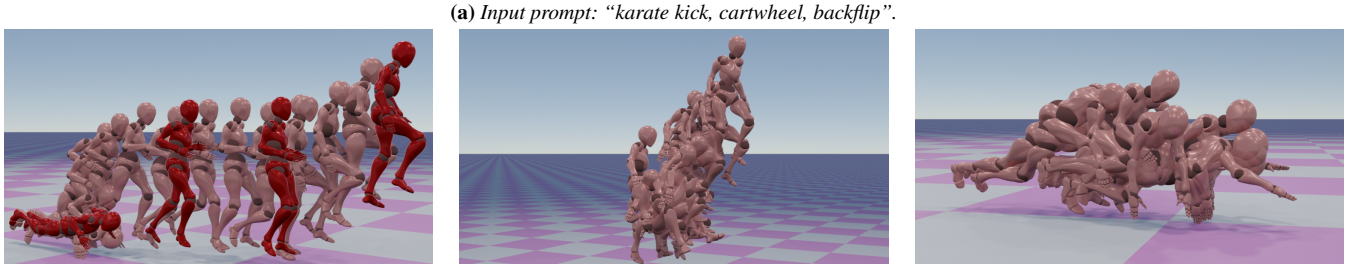
In contrast to other approaches that also regress the 6D continuous rotation representation of the 3D joints [ZBL*19], our model predicts the root-relative 3D joint positions and 4 additional variables that represent global information of the character, namely the *forward velocity*, *sideways velocity*, *rotation* around the *y*-axis, and character's *hip distance* w.r.t the floor. To properly compare our method in this evaluation, we calculate the joint angles by fitting a 3D body skeleton model through inverse kinematics (IK).

Even though our proposed text-to-motion approach uses a two-stage architecture with the main goal of supplementing the usability of the action-to-motion synthesis model, our approach is still competitive against other approaches specifically designed for the task of text-to-motion synthesis. Note that our results are computed after applying a post-process IK fitting which can further introduce additional errors and reduce our evaluation scores. The results are shown in Table 3.

4.3. Qualitative Results

In Fig. 6, we show synthesis results from our method rendered with a virtual character as well as the comparison against TEACH and

(a) *Input prompt: "karate kick, cartwheel, backflip".*



(b) *Input prompt: "lie down on the ground, jogging, walking up a stair".*

Figure 6: *Qualitative action-to-motion comparison between COMAND (left), TEACH (center), and MDM (Right).*

MDM. Our approach is able to synthesize very complex action combinations, even when such sequences were never observed in the training database. Moreover, our method performs exceptionally well at synthesizing the exact action prompt, i.e., it does not suffer from catastrophic neglect phenomena [CAV*23b], while existing approaches often fail, as can be observed from the action matching results in Tab. 2.

4.4. Limitations

Even though conceptually the backbone transformer architecture allows the model to generate a very long motion sequence, in practice, our result is limited by the maximum duration of the sequence observed in the training data. As a result, the quality of our model degrades if the target synthesis model is longer than 10 seconds. Exploring alternative seq-to-seq modeling solutions to mitigate this issue will be explored in future work.

5. Conclusion

We presented COMAND, a novel framework for multi-action 3D motion synthesis. Our method leverages an intermediate latent space that is intrinsically able to cluster a large-scale collection of motion sequences based on their natural and distinct actions, without requiring annotated action labels. The latent space can be used for action-to-motion synthesis, either by sampling the motion from the latent motion cluster, or by predicting the latent space from a given input. At the same time, the near-constant property of this proposed latent space makes it suitable as an intermediate representation for the text-to-motion synthesis task.

6. Acknowledgement

This work was supported by the ERC Consolidator Grant 4DRepLy (770784).

References

[AM19] AHUJA C., MORENCY L.-P.: Language2pose: Natural language grounded pose forecasting. In *International Conference on 3D Vision (3DV)* (2019), IEEE, pp. 719–728. 2, 6

[APBV22] ATHANASIOU N., PETROVICH M., BLACK M. J., VAROL G.: Teach: Temporal action composition for 3d humans. In *International Conference on 3D Vision (3DV)* (2022), IEEE. 5

[BRB*21] BHATTACHARYA U., REWKOWSKI N., BANERJEE A., GUHAN P., BERA A., MANOCHA D.: Text2gestures: A transformer-based network for generating emotive body gestures for virtual agents. In *2021 IEEE virtual reality and 3D user interfaces (VR)* (2021), IEEE, pp. 1–10. 6

[CAV*23a] CHEFER H., ALALUF Y., VINKER Y., WOLF L., COHEN-OR D.: Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. *arXiv preprint arXiv:2301.13826* (2023). 2

[CAV*23b] CHEFER H., ALALUF Y., VINKER Y., WOLF L., COHEN-OR D.: Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models, 2023. arXiv:2301.13826. 7

[CH05] CHAI J., HODGINS J. K.: Performance animation from low-dimensional control signals. *ACM Transactions on Graphics (TOG)* (2005). 2

[CJL*22] CHEN X., JIANG B., LIU W., HUANG Z., FU B., CHEN T., YU J., YU G.: Executing your commands via motion diffusion in latent space. *arXiv preprint arXiv:2212.04048* (2022). 6

[CMU03] CMU: CMU Graphics Lab Motion Capture Database, 2003. URL: http://mocap.cs.cmu.edu/. 2

[DMGT22] DABRAL R., MUGHAL M. H., GOLYANIK V., THEOBALT C.: Mofusion: A framework for denoising-diffusion-based motion synthesis. *arXiv preprint arXiv:2212.04495* (2022). 2, 6

[DN21] DHARIWAL P., NICHOL A.: Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems (NeurIPS) 34* (2021), 8780–8794. 2

[GCO*21] GHOSH A., CHEEMA N., OGUZ C., THEOBALT C., SLUSALLEK P.: Synthesis of compositional animations from textual descriptions. In *International Conference on Computer Vision (ICCV)* (2021), pp. 1396–1406. 2

[GMHP04] GROCHOW K., MARTIN S. L., HERTZMANN A., POPOVIĆ

Z.: Style-based inverse kinematics. *ACM Transactions on Graphics (TOG)* (2004). 2

[GZW*20] GUO C., ZUO X., WANG S., ZOU S., SUN Q., DENG A., GONG M., CHENG L.: Action2motion: Conditioned generation of 3d human motions. In *ACM International Conference on Multimedia* (2020), pp. 2021–2029. 2, 6

[GZWC22] GUO C., ZUO X., WANG S., CHENG L.: Tm2t: Stochastic and tokenized modeling for the reciprocal generation of 3d human motions and texts. In *European Conference on Computer Vision (ECCV)* (2022). 2, 6

[GZZ*22] GUO C., ZOU S., ZUO X., WANG S., JI W., LI X., CHENG L.: Generating diverse and natural 3d human motions from text. In *Computer Vision and Pattern Recognition (CVPR)* (2022), pp. 5152–5161. 2, 3, 4, 6

[HCS*22] HO J., CHAN W., SAHARIA C., WHANG J., GAO R., GRITSENKO A., KINGMA D. P., POOLE B., NOROUZI M., FLEET D. J., ET AL.: Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303* (2022). 2

[HJA20] HO J., JAIN A., ABBEEL P.: Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems (NeurIPS) 33* (2020), 6840–6851. 2

[HKPP20] HOLDEN D., KANOUN O., PEREPICHKA M., POPA T.: Learned motion matching. *ACM Transactions on Graphics (TOG)* (2020). 1

[HKS17] HOLDEN D., KOMURA T., SAITO J.: Phase-functioned neural networks for character control. *ACM Transactions on Graphics (TOG)* (2017). 1, 2, 3

[HSK16] HOLDEN D., SAITO J., KOMURA T.: A deep learning framework for character motion synthesis and editing. *ACM Transactions on Graphics (TOG)* (2016). 1, 2

[HSKJ15] HOLDEN D., SAITO J., KOMURA T., JOYCE T.: Learning motion manifolds with convolutional autoencoders. In *SIGGRAPH Asia 2015 Technical Briefs* (2015). 2

[KPH*21] KONG Z., PING W., HUANG J., ZHAO K., CATANZARO B.: Diffwave: A versatile diffusion model for audio synthesis. In *International Conference on Learning Representations (ICLR)* (2021). 2

[KW14] KINGMA D. P., WELLING M.: Auto-Encoding Variational Bayes. In *International Conference on Learning Representations (ICLR)* (2014). 3

[LGT*23] LIN C.-H., GAO J., TANG L., TAKIKAWA T., ZENG X., HUANG X., KREIS K., FIDLER S., LIU M.-Y., LIN T.-Y.: Magic3d: High-resolution text-to-3d content creation. In *Computer Vision and Pattern Recognition (CVPR)* (2023). 2

[LWB*10] LEE Y., WAMPLER K., BERNSTEIN G., POPOVIĆ J., POPOVIĆ Z.: Motion fields for interactive character locomotion. *ACM Transactions on Graphics (TOG)* (2010). 1

[LWH*23] LIU R., WU R., HOORICK B. V., TOKMAKOV P., ZAKHAROV S., VONDRICK C.: Zero-1-to-3: Zero-shot one image to 3d object, 2023. arXiv:2303.11328. 2

[MK05] MUKAI T., KURIYAMA S.: Geostatistical motion interpolation. *ACM Transactions on Graphics (TOG)* (2005). 2

[PBV21] PETROVICH M., BLACK M. J., VAROL G.: Action-conditioned 3d human motion synthesis with transformer vae. In *International Conference on Computer Vision (ICCV)* (2021), pp. 10985–10995. 2, 3, 5

[PBV22] PETROVICH M., BLACK M. J., VAROL G.: Temos: Generating diverse human motions from textual descriptions. In *European Conference on Computer Vision (ECCV)* (2022), pp. 480–497. 2, 3, 4, 5

[PCA*21] PUNNAKKAL A. R., CHANDRASEKARAN A., ATHANASIOU N., QUIROS-RAMIREZ A., BLACK M. J.: Babel: bodies, action and behavior with english labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2021), pp. 722–731. 5, 6

[PJBM22] POOLE B., JAIN A., BARRON J. T., MILDENHALL B.: Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988* (2022). 2

[RBL*22] ROMBACH R., BLATTMANN A., LORENZ D., ESSER P., OMMER B.: High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022), pp. 10684–10695. 2

[RDN*22] RAMESH A., DHARIWAL P., NICHOL A., CHU C., CHEN M.: Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125* (2022). 2

[RKH*21] RADFORD A., KIM J. W., HALLACY C., RAMESH A., GOH G., AGARWAL S., SASTRY G., ASKELL A., MISHKIN P., CLARK J., ET AL.: Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)* (2021), PMLR, pp. 8748–8763. 3

[RPG*21] RAMESH A., PAVLOV M., GOH G., GRAY S., VOSS C., RADFORD A., CHEN M., SUTSKEVER I.: Zero-shot text-to-image generation. In *International Conference on Machine Learning (ICML)* (2021), PMLR, pp. 8821–8831. 2

[SCS*22] SAHARIA C., CHAN W., SAXENA S., LI L., WHANG J., DENTON E. L., GHASEMIPOUR K., GONTIJO LOPES R., KARAGOL AYAN B., SALIMANS T., ET AL.: Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems (NeurIPS) 35* (2022), 36479–36494. 2

[SDCW19] SANH V., DEBUT L., CHAUMOND J., WOLF T.: Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108* (2019). 5

[SDWMG15] SOHL-DICKSTEIN J., WEISS E., MAHESWARANATHAN N., GANGULI S.: Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning (ICML)* (2015), PMLR, pp. 2256–2265. 2

[SMK22] STARKE S., MASON I., KOMURA T.: Deepphase: Periodic autoencoders for learning motion phase manifolds. *ACM Transactions on Graphics (TOG)* (2022). 2, 3

[SPH*22] SINGER U., POLYAK A., HAYES T., YIN X., AN J., ZHANG S., HU Q., YANG H., ASHUAL O., GAFNI O., ET AL.: Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792* (2022). 2

[TGH*22] TEVET G., GORDON B., HERTZ A., BERMANO A. H., COHEN-OR D.: Motionclip: Exposing human motion generation to clip space. In *European Conference on Computer Vision (ECCV)* (2022), pp. 358–374. 3

[TRG*22] TEVET G., RAAB S., GORDON B., SHAFIR Y., BERMANO A. H., COHEN-OR D.: Human motion diffusion model. *arXiv preprint arXiv:2209.14916* (2022). 2, 5, 6

[VdMH08] VAN DER MAATEN L., HINTON G.: Visualizing data using t-sne. *Journal of Machine Learning Research 9*, 11 (2008). 4

[ZBL*19] ZHOU Y., BARNES C., LU J., YANG J., LI H.: On the continuity of rotation representations in neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2019). 6

[ZCP*22] ZHANG M., CAI Z., PAN L., HONG F., GUO X., YANG L., LIU Z.: Motiondiffuse: Text-driven human motion generation with diffusion model. *arXiv preprint arXiv:2208.15001* (2022). 2, 6

[ZSKS18] ZHANG H., STARKE S., KOMURA T., SAITO J.: Mode-adaptive neural networks for quadruped motion control. *ACM Transactions on Graphics (TOG)* (2018). 1

[ZZC*23] ZHANG J., ZHANG Y., CUN X., HUANG S., ZHANG Y., ZHAO H., LU H., SHEN X.: T2m-gpt: Generating human motion from textual descriptions with discrete representations. *arXiv preprint arXiv:2301.06052* (2023). 2, 6