Sequential Texts Driven Cohesive Motions Synthesis with Natural Transitions

Shuai Li[1,3], Sisi Zhuang[1], Wenfeng Song[2*], Xinyu Zhang[2], Hejia Chen[1], Aimin Hao[1]
[1]State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, P.R. China
[2]Computer School, Beijing Information Science and Technology University, P.R. China
[3]Zhongguancun Laboratory, Beijing, P.R. China
{lishuai, sisizhuang}@buaa.edu.cn, songwenfenga@gmail.com
zhangxinyu1@bistu.edu.cn, {chenhj2000, ham}@buaa.edu.cn

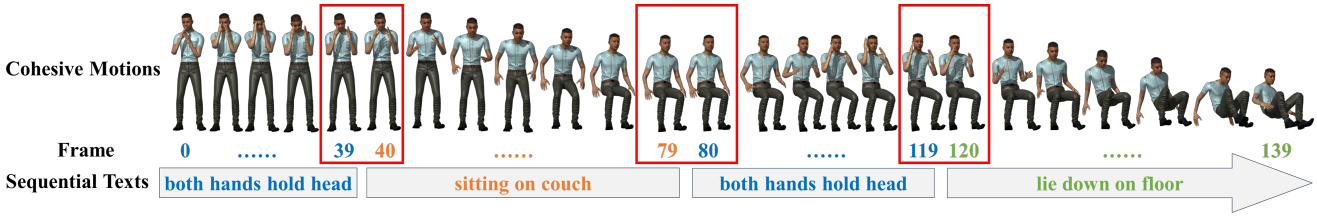

Figure 1: Our method synthesizes semantic human motion with natural transition using free-form sequential texts. The skeleton motion can be easily transferred to any character. The character used for demonstration is from Mixamo [1].

Abstract

The intelligent synthesis/generation of daily-life motion sequences is fundamental and urgently needed for many VR/metaverse-related applications. However, existing approaches commonly focus on monotonic motion generation (e.g., walking, jumping, etc.) based on single instruction-like text, which is still not intelligent enough and can't meet practical demands. To this end, we propose a cohesive human motion sequence synthesis framework based on free-form sequential texts while ensuring semantic connection and natural transitions between adjacent motions. At the technical level, we explore the local-to-global semantic features of previous and current texts to extract relevant information. This information is used to guide the framework in understanding the semantics of the current moment. Moreover, we propose learnable tokens to adaptively learn the influence range of the previous motions towards natural transitions. These tokens can be trained to encode the relevant information into well-designed transition loss. To demonstrate the efficacy of our method, we conduct extensive experiments and comprehensive evaluations on the public dataset as well as a new dataset produced by us. All the experiments confirm that our method outperforms the state-of-the-art methods in terms of semantic matching, realism, and transition fluency. Our project is public available. https://druthrie.github.io/sequential-texts-to-motion/

*Corresponding author

1. Introduction and Motivation

Human motion synthesis is fundamental for numerous application, especially for virtual reality, games, and metaverse-related applications [10, 19, 5, 3, 36, 15, 22, 20, 24], of which, it is in high demand to accurately control the digital human motion with natural language. Existing approaches commonly focus on monotonic motion synthesis/generation based on single instruction-like text description. However, practical applications usually require digital humans to respond to multiple rounds of sustainable interactions, wherein they can continuously generate reasonable responses to the sequential texts. Therefore, given a set of free-form sequential texts, we aim to synthesize a cohesive human motion sequence. Namely, the animation clips should be consistent with the ongoing text descriptions, and the whole motion sequence should have smooth semantic connection and natural transitions.

At present, sequential texts-driven motion synthesis has not been well studied, mainly due to lacking long-term continuous motion datasets with accompanying free-form text descriptions. Recently, TEACH [6] first attempts to address this problem by proposing a dataset (we refer to it as BABEL-TEACH). Each item in the BABEL-TEACH dataset contains two adjacent text-motion pairs. However, TEACH [6] builds on the one-time method TEMOS [29], wherein the limited 5-frame motion is added to encode together. However, it does not consider the potential benefits of previous text information. In practice, previous texts can provide extra and more accurate semantic information for

current motion synthesis. Therefore, global semantic needs to be extracted from the previous text and the local semantics from the current text, and then fuses the semantics to guide the semantic understanding at each moment.

Besides, there are still two main challenges. Firstly, the semantic relationship between adjacent texts is largely overlooked in previous research. For example, when synthesizing a motion for the current text "touching the face with left hand," the previous text context should be considered. Specifically, if the previous text was "sitting on a chair," the synthesized motion should depict a person sitting and touching face with left hand (shown in Fig. 2); if the previous text was "a person kicks a ball with right foot," the motion should show a standing person touching face with left hand. This requires a more comprehensive understanding about the context and semantic relationships between the texts. Secondly, existing methods tend to abrupt transitions between adjacent motions when multiple motions are synthesized separately and stitched together. We aim to address this issue by seamlessly blending the synthesized motions. It should intelligently capture the temporal and spatial continuity between the motions to ensure a more natural and coherent sequence of motions.

Considering the easy scalability of autoregressive methods, we should extend previous autoregressive single text-driven motion synthesis methods to multiple motions involved in motion sequence synthesis. We observe that a similar human body posture can be achieved by transferring the end features of the previous motion to the next motion, although this is not always entirely consistent. This suggests that previous motion information should be leveraged more effectively and sufficiently rather than being adopted straightforwardly. Therefore, we plan to introduce a transition reasoning module that adaptively learns attention score from the previous motion information and infers accurate motion features at the transition, making the transition between adjacent motions more natural and smooth.

To this end, we create a new dataset with much longer-term motions (2-5 times longer) based on a synthetic method, and the corresponding text descriptions are more abundant. We conduct experiments on both BABEL-TEACH [6] dataset and our sequential texts described motion (STDM) dataset. The results demonstrate that our method outperforms existing methods in terms of semantic matching, realism, and transition fluency. Specially, the salient contributions can be summarized as follows.

- We propose a cohesive motions synthesis framework using sequential texts as inputs. The framework can integrate previous text information to obtain semantic features and adaptively select valid previous motion information to guide the current motion synthesis.
- We design a local-to-global (L2G) semantic fusion module to extract accurate contextual information. It can well take into account global-context information from the previous text to infer the current moment's text meaning to guide the semantics-consistent motion synthesis.
- We introduce a transition reasoning module to adaptively select motion snippet from previous motion information to make the synthesized motion natural and smooth w.r.t the previous, wherein a well-defined transition loss is employed to further constrain the fluency of the transitions.
- We create a new sequential texts described motion (STDM) dataset, which involves more extended motion frames and more diverse text descriptions than the existing dataset.

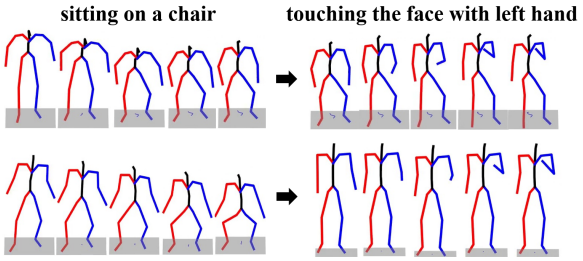

Figure 2: **Top row:** Motion synthesis corresponding to sequential texts by our method. **Bottom row:** Motion synthesis corresponding to single text by T2M [11].

2. Related Work

We briefly introduce the related works of human motion synthesis considering semantic control, including single text driven motion synthesis and sequential texts driven motion synthesis.

Single text driven motion synthesis. The motion synthesis from text began with the synthesized motion from action labels, but this approach was limited in covering most human motions [21, 27, 35, 38, 28, 12, 28, 7]. Later, researchers [4, 9, 34] focused on synthesizing motion from complex free-form text descriptions, with early work generating single, deterministic motions from text. Recent works have aimed to improve the diversity and details of the synthesized motion. For example, TEMOS [29] and T2M [11] utilized VAE architecture to synthesize diverse motions from the same texts. The main difference between them is that the former utilized Transformer architecture to synthesize all moment motions at once, while the latter utilized GRU architecture to synthesize every moment motion auto-regressively. To capture more detailed information in the text, TM2T [14] utilized VQVAE-based motion markers to provide a fair environment when considering motion and text signals, and utilized motion-to-text analysis module to strengthen the constraints on text-to-motion synthesis. With the successful application of the diffusion model in image generation, MotionDiffuse [37] utilized the diffusion model

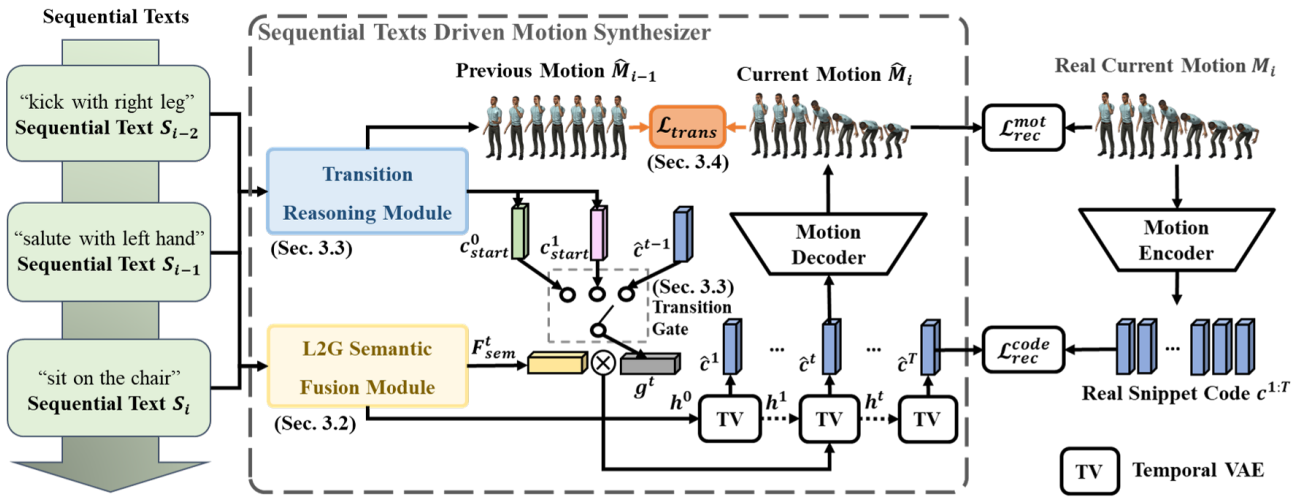

Figure 3: **Method Overview:** Our model takes sequential texts as input. In the local-to-global (L2G) semantic fusion module, the previous text's global features and the current text's local features are fused and extracted. In the transition reasoning module, the motion snippet code at the transition of adjacent motion ($\mathbf{c}^0_{start}$ and $\mathbf{c}^1_{start}$) is deduced from the previous motion. Input the gate snippet code $\mathbf{g}^t$ and the semantic text feature F^t_{sem} to synthesize the motion snippet code $\hat{\mathbf{c}}^t_i$. The transition loss further limits the fluency of the transition with the previous motion.

to simulate text driven conditional human motion synthesis, which can better respond to fine-grained instructions of body parts. FLAME [18] utilized transformer-based diffusion model architecture to synthesize human motion, and it can allow editing frames and joints without fine-tuning. Its editing ability can be extended to motion prediction or in-between tasks. Although the diffusion model improves the quality of the synthesized motion, its inference speed is relatively slow, so it is unsuitable for real-time applications.

Sequential texts driven motion synthesis. Synthesizing motion sequences from sequential texts is challenging due to the limited availability of long-term continuous motions datasets with text descriptions. Previous works Action2video [13] realized the transition between actions by changing the input action label type in an autoregressive synthesizer. Yet, they only conducted experiments on a limited number of action labels in three datasets: NTU-RGBD [33], CMU [2], and humanACT12 [12]. Mao et al. [26] proposed a weakly-supervised action-driven motion prediction method, but they only used 20 action categories with clear transitions in BABEL [32] to synthesize the following motion and transition. Based on the existing BABEL dataset [32], TEACH [6] proposed a new dataset, in which each data consisted of two adjacent text descriptions and motion sequences. The model of TEACH [6] encoded five previous motion frames using TEMOS [29], every single motion sequence was synthesized simultaneously, and adjacent motion sequences were auto-regressively synthesized. Our work builds on this by proposing a cohesive human motion sequence synthesis framework based on free-form sequential texts, ensuring semantic connection and natural transitions between adjacent motions.

3. Our Approach

3.1. Method Overview

The overall pipeline of the proposed model is shown in Fig. 3. The temporal VAE and motion autoencoder used in the model are similar to those used in the T2M model, wherein it is employed to encode the motion sequence into a motion snippet code sequence and reconstruct the motion sequence with a decoder. To take sequential texts as input and ensure semantically coherent synthesized motions, a local-to-global (L2G) semantic fusion module is introduced (Sec. 3.2). This module guides learning accurate semantic features from the input texts at each moment, including historical text information. To ensure smooth and natural transitions between adjacent motions, a transition reasoning module is introduced (Sec. 3.3). This module dynamically deduces the motion snippet code during transition based on previous motion information, which can help the model generate realistic and coherent transitions. Additionally, a transition loss is proposed to enhance the smoothness of the transition motion (Sec. 3.4). Overall, the proposed pipeline combines several modules and techniques to generate realistic and semantically coherent motion sequences from sequential texts input. The model can generate smooth transitions between motions, and can adaptively learn the influence range of previous motions to further improve the quality of the synthesized motion sequences.

Preliminaries. Our method aims to take free-form sequential texts $\mathbf{S} = (\mathbf{S}_1, \mathbf{S}_2, \cdots, \mathbf{S}_n)$ as input, e.g., $\mathbf{S}$=(*"salute with the left hand", "salute with the right hand", "sit on the chair", "crawl"*), and outputs corresponding human motion sequences $\mathbf{M} =$

$(\mathbf{M}_1, \mathbf{M}_2, \cdots, \mathbf{M}_n)$. The transition between adjacent motions should be natural and smooth, and the synthesized motion should reflect the semantic information of previous texts. The data format is consistent with HumanML3D [11] dataset, including the root angular velocity along the Y-axis, the root linear velocity on the XZ-plane, the root height, the local joints positions, velocities, and rotations [40] in root space, while the motions follow the skeleton structure of SMPL [25] with 22 joints. Each moment in our model represents 4 frames.

3.2. Local-to-global Semantic Fusion Module

To synthesize motion that accurately matches the semantics of sequential texts, our model relies on accurately extracting text features at each moment to guide the synthesis process. As the motion we synthesize is based on sequential texts, the current motion is determined not only by the current text semantics but also by previous text semantics. To capture the temporal clues, we encode both the previous and current text, and use an attention mechanism to extract text features that integrate the previous text semantics at each moment (shown in Fig. 4). For previous texts, we use a pre-trained CLIP [31] text encoder to extract global features of previous texts. In order to make the features more suitable for our task in the process of training, linear projector is used to extract features further. The global features from the previous texts are expressed as:

$$\mathbf{G}_{pre} = f(CLIP_{text}(\mathbf{S}_{i-1})). \tag{1}$$

It is difficult for the CLIP text encoder to extract the local features of the text, motion synthesis of each moment requires local text semantic guidance, so we employ a Bi-GRU-based text encoder, similar to T2M [11], to obtain the local features $\mathbf{L}_{1:m}$ from the current text. To capture accurate text semantic features at the arbitrary frames, we use an attention mechanism, wherein the key $\mathbf{K}(L)$ and value $\mathbf{V}(L)$ of attention are the local features of the current text. Since the semantic information of the previous text is required for guidance at the beginning of the currently synthesizing motion, we use the global features of the previous text as the initial query $\mathbf{Q}(G)$. At subsequent moments, the motion snippet code generator in temporal VAE provides the queries $\mathbf{Q}(L)$. This can be expressed mathematically as:

$$\begin{aligned} \mathbf{Q}(G) = \mathbf{G}_{pre}\mathbf{W}^Q, \mathbf{Q}(L) = \mathbf{h}^{t-1}\mathbf{W}^Q, \\ \mathbf{K}(L) = \mathbf{L}_{1:m}\mathbf{W}^K, \mathbf{V}(L) = \mathbf{L}_{1:m}\mathbf{W}^V, \\ \mathbf{F}^t_{sem} = f(\frac{\mathbf{Q}(G/L)\mathbf{K}(L)^T}{\sqrt{d_{sem}}})\mathbf{V}(L), \end{aligned} \tag{2}$$

where $\mathbf{W}^k, \mathbf{W}^V \in \mathbb{R}^{d_w \times d_{sem}}$ and $\mathbf{W}^Q \in \mathbb{R}^{d_h \times d_{sem}}$ are trainable weights;d_h, d_w and d_{sem} are the number of channels in hidden unit $\mathbf{h}^{t-1}$, current text feature $\mathbf{L}^{1:m}$ and semantic fusion attention layer, respectively. $\mathbf{h}^t$ is the hidden unit generated by the motion snippet code generator at each moment and $\mathbf{h}^0$ is obtained from $\mathbf{G}_{pre}$. $\mathbf{F}^t_{sem}$ is the text semantic feature at moment t.

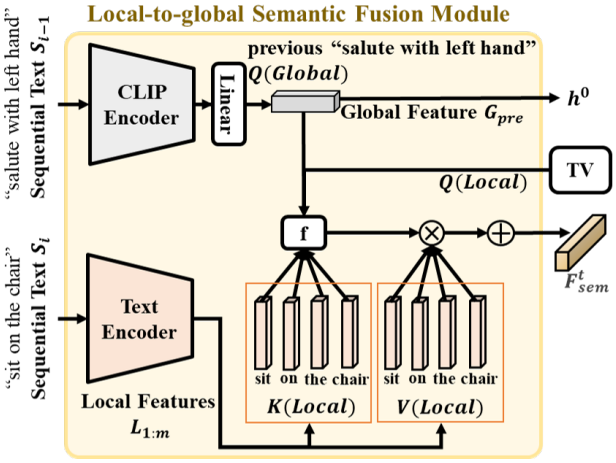

Figure 4: **Local-to-global semantic fusion module:** After extracting the global feature $\mathbf{G}_{pre}$ from the previous texts and the local features $\mathbf{L}_{1:m}$ from the current text, feature fusion is performed to obtain the accurate semantic feature $\mathbf{F}^t_{sem}$ of the current text at moment t.

3.3. Transition Reasoning Module

When synthesizing a new motion, the starting pose is influenced by the ending pose of the previous motion. Grafting the end motion snippet code of the previous motion to the start moment of the current motion can produce similar human poses during the transition. However, the transition between the two motions is still unnatural and abrupt. To address this, we introduce a learnable start motion snippet code token $\mathbf{c}_{token}$ and utilize it along with the previous motion snippet code sequence in a transformer encoder. The output of the token position is used as the start motion snippet code $\mathbf{c}_{start}$ for the current motion, allowing the model to learn the range and weight of the previous motions (shown in Fig. 5). Additionally, the length of the current motion affected by the previous motion information needs to be determined. Through experimentation, we obverse that the best performance is achieved by reasoning the first two motion snippet codes ($\mathbf{c}^0_{start}$ and $\mathbf{c}^1_{start}$) of the current motion from the previous motions. The verification experiment is in Sec. 4.5. The mathematical expression is as:

$$\mathbf{c}^0_{start}, \mathbf{c}^1_{start} = Transformer(\hat{\mathbf{c}}_{i-1}, \mathbf{c}^0_{token}, \mathbf{c}^1_{token}). \tag{3}$$

As shown by the transition gate in Fig. 3, when synthesizing the motion snippet codes of the first two moments, we choose the motion snippet codes ($\mathbf{c}^0_{start}$ and $\mathbf{c}^1_{start}$) inferred from the previous motion as gate snippet code $\mathbf{g}^t$. From the third moment, the motion snippet code $\hat{\mathbf{c}}^{t-1}$ from the previ-

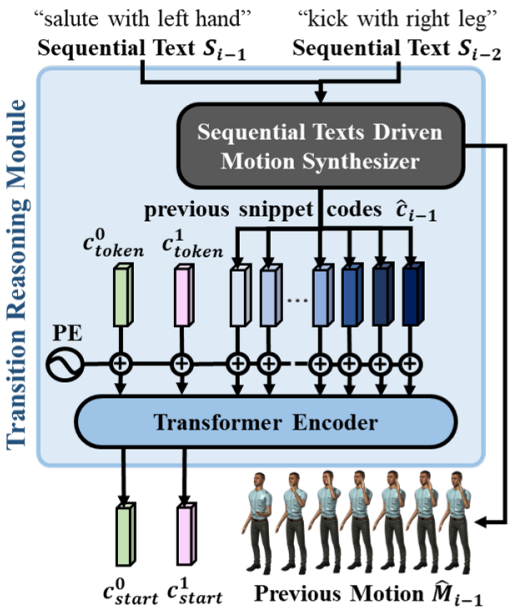

Figure 5: **Transition reasoning module:** Given the previous texts to get the previous synthesized motion sequence $\hat{\mathbf{M}}_{i-1}$ and motion snippet codes $\hat{\mathbf{c}}_{i-1}$, we use two learnable tokens ($\mathbf{c}^0_{token}$ and $\mathbf{c}^1_{token}$) to learn the previous motion information that needs attention, so as to obtain the motion snippet codes at the starting moment of the current motion ($\mathbf{c}^0_{start}$ and $\mathbf{c}^1_{start}$).

ous moment is used, which is synthesized autoregressively. For gate snippet code $\mathbf{g}^t$, the mathematical expression is as:

$$\mathbf{g}^t = \begin{cases} \mathbf{c}^0_{start} & t = 1 \\ \mathbf{c}^1_{start} & t = 2 \\ \hat{\mathbf{c}}^{t-1} & t \geq 3 \end{cases} \tag{4}$$

When synthesizing the motion based on the first text $\mathbf{S}_1$ of the sequential texts, we set the previous texts as the default text $\mathbf{S}_0$=*"start"*, and the starting snippet code $\mathbf{c}^0_{start}$ is obtained from the mean pose.

3.4. Transition Loss & Total Loss

We conduct a statistical analysis of the dataset and find that the motion gap (L1 distance) of almost all adjacent frames is less than 0.1. Inspired by this, we design a transition loss L_{Trans} to reduce the motion gap between the adjacent frames and avoid abrupt transition. The representation of the L_{Trans} is

$$\mathcal{L}_{Trans} = \|\hat{\mathbf{M}}^1_i - \hat{\mathbf{M}}^{T_{i-1}}_{i-1}\|_1. \tag{5}$$

The final loss function of our model consists of 4 parts: motion snippet code reconstruction loss $\mathcal{L}^{code}_{rec}$, motion reconstruction loss $\mathcal{L}^{mot}_{rec}$, KL loss $\mathcal{L}_{KL}$ of prior distribution and posterior distribution in temporal VAE, and the transition loss $\mathcal{L}_{Trans}$. The $\lambda_{code}, \lambda_{mot}, \lambda_{KL}, \lambda_{Trans}$ are the corresponding weights of the losses.

$$\mathcal{L} = \lambda_{code}\mathcal{L}^{code}_{rec} + \lambda_{mot}\mathcal{L}^{mot}_{rec} + \lambda_{KL}\mathcal{L}_{KL} + \lambda_{Trans}\mathcal{L}_{Trans}. \tag{6}$$

Dataset	Quantity	Duration	Vocab.
STDM	5289	20.02 h	2455
BABEL-TEACH[6]	16266	21.86 h	1037

Table 1: **Comparison of datasets:** STDM dataset has longer-term frames within a single motion, and the text is more abundant.

4. Experiments

First, we introduce our STDM dataset (Sec. 4.1), evaluation metrics (Sec. 4.2), and then conduct quantitative (Sec. 4.3) and qualitative (Sec. 4.6) evaluations. We carry out an ablation study (Sec. 4.4) to prove the effectiveness of our proposed modules and analyze the parameters (Sec. 4.5).

4.1. STDM Dataset

Each text-motion pair in most existing human motion datasets [30, 17, 23, 39, 12, 8] is independent of the other, and does not contain adjacent text-motion pairs. In a recent study, TEACH [6] used the BABEL [32] dataset to obtain a BABEL-TEACH dataset containing two adjacent text-motion pairs, which is consistent with our research goal.

To obtain adequate motion sequences, we synthesize sequential texts and motion sequences by modifying the existing single text driven motion synthesis method from the viewpoint of data generation instead of using a mocap device. By randomly selecting and annotating texts from a large pool, we create sequential text fed into our improved T2M [11], generating motion sequences. Based on T2M [11], we find that the motion sequences decoded by the same motion snippet code are similar. Our approach diverges from T2M: In the inference stage, instead of using a fixed mean pose, we use snippet codes from the last eight frames of the previous motion, ensuring smoother and more natural transitions. Our experiment results show that the motion sequences synthesized in this way are similar in transition. Since the motion generated by the generation model may not match the semantics of the input text and the transition may not be smooth, we manually remove data with poor transitions and correct erroneous texts to maintain the quality of our STDM dataset. After filtering, we get 5289 pairs containing two adjacent text-motion. Our dataset is collected under the same structure as HumanML3D [11].

Compared with the BABEL-TEACH [6] dataset, STDM has longer-term motion frames, with a single motion sequence ranging from 2 to 10 seconds, while most BABEL-TEACH [6] data is less than 3 seconds. However, the text

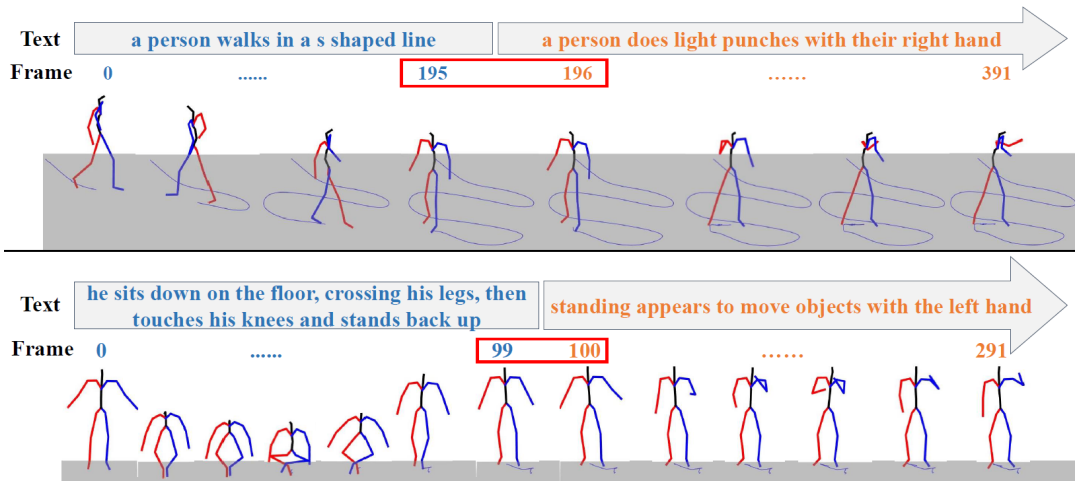

Figure 6: **STDM Dataset:** Two examples in the STDM dataset, which involves more extended motion frames and more diverse text descriptions than the BABEL-TEACH [6] dataset. Blue traces represent the digital human's trajectory.

descriptions in BABEL-TEACH [6] usually describe only one action, such as "walk forward". The text descriptions in the STDM are more abundant, such as "a person raises his right hand and walks forward". A detailed comparison of the two datasets can be found in Tab. 1. Fig. 6 shows two examples of our STDM dataset.

4.2. Evaluation Metrics

In order to evaluate the quality of the synthesized motion sequence, we use the feature extractor proposed by Guo et al. [11] to extract the features of text and motion. This feature extractor is trained with contrast loss [16], which can make the matched text-motion pairs produce similar feature vectors. Therefore, the feature extractor can measure whether the text and motion sequence match. We combine two adjacent text-motion pairs into one to train the feature extractor. We use the same evaluation metrics as T2M [11]. The metrics are described in the supplementary materials.

To demonstrate the fluency of different methods when transiting between adjacent motions, we add a metric to evaluate the fluency of transition, by calculating the distance between adjacent frames of two adjacent motion sequences. For transition fluency, the mathematical expression is as:

$$F = \sum_{i=1}^{N} \|\mathbf{M}_{i1}^{T1} - \mathbf{M}_{i2}^{1}\|_1. \tag{7}$$

Here, we calculate the average transition distance (L_1 distance) of the motion sequence synthesized by N adjacent text pairs. For the i-th adjacent text pair, the synthesized adjacent motion sequence pair is $\mathbf{M}_{i1} = (\mathbf{M}_{i1}^{1}, \mathbf{M}_{i1}^{2}, ..., \mathbf{M}_{i1}^{T1})$ and $\mathbf{M}_{i2} = (\mathbf{M}_{i2}^{1}, \mathbf{M}_{i2}^{2}, ..., \mathbf{M}_{i2}^{T2})$.

4.3. Quantitative Evaluation

Baselines. We compare our method with T2M [11], our two variants T2M-Joint and Complusion-Code, and TEACH [6]. For T2M, we input a single text, output a single motion, and align the root coordinates of the last frame of the previous motion with the root coordinates of the first frame of the next motion. For T2M-Joint, we keep the T2M model structure and train the model by combining two adjacent texts with commas as input. For Complusion-Code, based on the T2M [11], the last motion snippet code of the previous motion is used as the start motion snippet code of the next motion during training (Unlike the STDM dataset generation method, the dataset generation only uses the snippet codes of previous motions in the inference phase. Complusion-Code also uses the snippet codes of previous motions in the training phase). For TEACH [6], we compared two versions using spherical linear interpolation (Slerp) and not using Slerp, respectively. Slerp operation is unrelated to the TEACH [6] model and is simply a way of post-processing synthesized data.

Tab. 2 and Tab. 3 show our quantitative comparison results with other baselines on BABEL-TEACH [6] and STDM datasets. We divide the training, testing, and validation sets according to the ratio of 0.8 : 0.15 : 0.05. For a fair comparison, we conduct 20 experiments and show the values with confidence in the 95% range. From the results, our method outperforms all other methods, especially in the transition fluency metric, which is important for synthesizing natural and smooth transition motion. T2M [11] can only synthesize a single motion per inference, and there is no connection between adjacent motions, so it performs poorly in all indicators. Because our testing data are composed of two adjacent texts, the T2M-Joint method performs similarly to the real data regarding transition fluency

Methods	R Precision↑			FID↓	MultiModal Dist↓	Transition Fluency↓
	Top 1	Top 2	Top 3			
Real motions	$0.560^{\pm 0.002}$	$0.750^{\pm 0.003}$	$0.834^{\pm 0.002}$	$0.001^{\pm 0.000}$	$3.448^{\pm 0.001}$	-
T2M [11]	$0.376^{\pm 0.003}$	$0.543^{\pm 0.002}$	$0.647^{\pm 0.003}$	$6.890^{\pm 0.081}$	$5.218^{\pm 0.010}$	$1.108^{\pm 0.004}$
T2M-Joint	$0.373^{\pm 0.004}$	$0.534^{\pm 0.003}$	$0.640^{\pm 0.003}$	$2.727^{\pm 0.032}$	$4.983^{\pm 0.011}$	-
Compulsion-Code	$0.432^{\pm 0.003}$	$0.613^{\pm 0.004}$	$0.717^{\pm 0.003}$	$3.726^{\pm 0.068}$	$4.612^{\pm 0.023}$	$0.348^{\pm 0.002}$
TEACH (no Slerp) [6]	$\mathbf{0.573}^{\pm 0.002}$	$\mathbf{0.756}^{\pm 0.003}$	$\mathbf{0.842}^{\pm 0.002}$	$2.263^{\pm 0.055}$	$\mathbf{3.487}^{\pm 0.010}$	$0.577^{\pm 0.003}$
TEACH (Slerp) [6]	$\underline{0.563}^{\pm 0.003}$	$\underline{0.750}^{\pm 0.003}$	$\underline{0.839}^{\pm 0.003}$	$\underline{2.240}^{\pm 0.053}$	$\underline{3.526}^{\pm 0.011}$	$\underline{0.118}^{\pm 0.001}$
Ours	$\underline{0.542}^{\pm 0.003}$	$\underline{0.728}^{\pm 0.003}$	$\underline{0.818}^{\pm 0.003}$	$\underline{1.628}^{\pm 0.031}$	$\underline{3.662}^{\pm 0.010}$	$\underline{0.177}^{\pm 0.001}$
Ours (Slerp)	$0.519^{\pm 0.004}$	$0.705^{\pm 0.004}$	$0.803^{\pm 0.003}$	$\mathbf{1.548}^{\pm 0.024}$	$3.797^{\pm 0.012}$	$\mathbf{0.036}^{\pm 0.000}$

Table 2: **Quantitative evaluation on testing data of BABEL-TEACH:** $\pm$ indicates 95% confidence interval, $\uparrow$ and $\downarrow$ respectively denotes better performance with larger or lower value. **Bold** indicates the optimal result, the underscore represents the suboptimal result, while wave line refers to the third best. The results show that the motion synthesized by our model outperforms other baselines in terms of semantic matching, transition fluency, and realism.

Methods	R Precision↑			FID↓	MultiModal Dist↓	Transition Fluency↓
	Top 1	Top 2	Top 3			
Real motions	$0.322^{\pm 0.004}$	$0.505^{\pm 0.005}$	$0.625^{\pm 0.006}$	$0.014^{\pm 0.002}$	$3.461^{\pm 0.005}$	$0.454^{\pm 0.000}$
T2M [11]	$0.322^{\pm 0.006}$	$0.493^{\pm 0.005}$	$0.614^{\pm 0.005}$	$1.389^{\pm 0.049}$	$3.500^{\pm 0.009}$	$0.639^{\pm 0.003}$
T2M-Joint	$0.304^{\pm 0.005}$	$0.488^{\pm 0.006}$	$0.613^{\pm 0.005}$	$\mathbf{0.744}^{\pm 0.042}$	$3.492^{\pm 0.011}$	-
Compulsion-Code	$0.285^{\pm 0.007}$	$0.450^{\pm 0.006}$	$0.572^{\pm 0.006}$	$3.146^{\pm 0.080}$	$3.806^{\pm 0.017}$	$0.330^{\pm 0.003}$
TEACH (no Slerp) [6]	$0.318^{\pm 0.006}$	$\underline{0.505}^{\pm 0.005}$	$\underline{0.634}^{\pm 0.005}$	$1.414^{\pm 0.055}$	$\underline{3.483}^{\pm 0.012}$	$0.246^{\pm 0.002}$
TEACH (Slerp) [6]	$\underline{0.326}^{\pm 0.004}$	$0.504^{\pm 0.005}$	$0.631^{\pm 0.006}$	$1.416^{\pm 0.054}$	$3.487^{\pm 0.013}$	$\underline{0.049}^{\pm 0.000}$
Ours	$\mathbf{0.328}^{\pm 0.006}$	$\underline{0.510}^{\pm 0.006}$	$\underline{0.633}^{\pm 0.005}$	$\underline{1.085}^{\pm 0.063}$	$\underline{3.441}^{\pm 0.017}$	$\underline{0.109}^{\pm 0.001}$
Ours (Slerp)	$\underline{0.325}^{\pm 0.007}$	$\mathbf{0.515}^{\pm 0.007}$	$\mathbf{0.637}^{\pm 0.005}$	$\underline{1.134}^{\pm 0.059}$	$\mathbf{3.440}^{\pm 0.013}$	$\mathbf{0.022}^{\pm 0.000}$

Table 3: **Quantitative evaluation on the testing data of STDM:** The results show that the motion synthesized by our model outperforms other baselines in terms of semantic matching, transition fluency, and realism.

(less than 0.1). However, when the input text consists of more than two sentences, the T2M-Joint method cannot guarantee transition fluency. Because Compulsion-Code uses the end motion snippet code of the previous motion as the start snippet code of the next motion, the performance of transition fluency is better than T2M [11], but it is still not as good as ours. Although TEACH [6] is slightly better than ours in R Precision and MultiModal Distance on the BABEL-TEACH [6] dataset, as shown in Tab. 2, its transition fluency is not as good as ours, we can obtain transitional and natural motion sequences without Slerp operation, so our method is better than TEACH [6] in a comprehensive evaluation. Although the synthetic STDM dataset is less natural than actual mocap data in transition fluency, we find that our method can compensate for the defects of the training data after training. In terms of transition fluency, the results synthesized by our method outperform the ground truth of the testing data. This further verifies the effectiveness of our method in synthesizing natural and smooth transition motions.

User study. We randomly select 50 texts from the testing data and arrange 3 to 5 texts together to form a set of sequential texts. Some text combinations are not present in the dataset to test the motion generation effect of each model for text combinations not in the dataset. The synthesized motions based on these sequential text groups are randomly shuffled and presented using different methods. Users are required to score synthetic motions from the following three aspects: (1) Matching degree: the degree of semantic matching between the motion and the text; (2) Transition fluency: the natural degree of transition between adjacent motions; (3) Realism: how realistic the motion is. The result is shown in Fig. 7.

4.4. Ablation Study

In order to illustrate the influence of the modules and training strategies we introduced on the overall model, we carried out the ablation experiment. The results are shown in Tab. 4.

The first experiment is to remove the transition reasoning module. We set the skeleton to mean pose to initialize the motion snippet code at the start moment. However, the model does not know the human posture at the beginning, so it greatly reduces transition fluency (value increased by 0.417). The experiment results show that the performance of the model will decline without the transition reasoning module because it does not aware of the information of previous motion when the model synthesizes the motion sequence of the current text, even though the model can synthesize motion sequences that match the semantics of the

Methods	R Precision↑			FID↓	MultiModal Dist↓	Transition Fluency↓
	Top 1	Top 2	Top 3			
Real motions	$0.561^{\pm0.003}$	$0.747^{\pm0.003}$	$0.836^{\pm0.002}$	$0.002^{\pm0.000}$	$3.449^{\pm0.002}$	-
w/o Transition reasoning module	$0.535^{\pm0.004}$	$0.713^{\pm0.003}$	$0.805^{\pm0.003}$	$2.190^{\pm0.038}$	$3.719^{\pm0.013}$	$0.594^{\pm0.003}$
w/o L2G semantic fusion module	$0.440^{\pm0.003}$	$0.621^{\pm0.003}$	$0.720^{\pm0.003}$	$4.592^{\pm0.071}$	$4.593^{\pm0.013}$	$\mathbf{0.151}^{\pm0.001}$
w/o Transition loss	$0.525^{\pm0.004}$	$0.706^{\pm0.004}$	$0.801^{\pm0.002}$	$1.766^{\pm0.036}$	$3.797^{\pm0.011}$	$0.293^{\pm0.001}$
w/o Segmented training strategy	$0.500^{\pm0.002}$	$0.684^{\pm0.004}$	$0.783^{\pm0.004}$	$2.164^{\pm0.038}$	$4.041^{\pm0.013}$	$0.328^{\pm0.002}$
Ours	$\mathbf{0.542}^{\pm0.003}$	$\mathbf{0.726}^{\pm0.003}$	$\mathbf{0.817}^{\pm0.002}$	$\mathbf{1.588}^{\pm0.030}$	$\mathbf{3.655}^{\pm0.013}$	$\underline{0.177}^{\pm0.001}$

Table 4: **Ablation study:** The results show that the transition reasoning module has the most significant impact on transition fluency, the L2G semantic fusion module has the most significant impact on R Precision, FID, and MultiModel Distance, which represent semantic matching and realism, and the transition loss and segmentation training strategy have a certain influence on all indicators.

Numbers of snippet code	R Precision↑			FID↓	MultiModal Dist↓	Transition Fluency↓
	Top 1	Top 2	Top 3			
Real motions	$0.561^{\pm0.003}$	$0.750^{\pm0.002}$	$0.834^{\pm0.002}$	$0.002^{\pm0.000}$	$3.449^{\pm0.001}$	-
1 snippet code	$0.517^{\pm0.003}$	$0.693^{\pm0.004}$	$0.786^{\pm0.004}$	$2.027^{\pm0.043}$	$3.845^{\pm0.014}$	$\mathbf{0.173}^{\pm0.001}$
2 snippet codes	$\mathbf{0.540}^{\pm0.003}$	$\mathbf{0.725}^{\pm0.003}$	$\mathbf{0.816}^{\pm0.002}$	$\mathbf{1.602}^{\pm0.026}$	$\mathbf{3.662}^{\pm0.011}$	$0.177^{\pm0.001}$
3 snippet codes	$0.522^{\pm0.004}$	$0.702^{\pm0.003}$	$0.796^{\pm0.002}$	$2.050^{\pm0.038}$	$3.838^{\pm0.015}$	$0.175^{\pm0.001}$
5 snippet codes	$0.504^{\pm0.003}$	$0.687^{\pm0.004}$	$0.785^{\pm0.003}$	$1.781^{\pm0.029}$	$3.913^{\pm0.013}$	$0.213^{\pm0.001}$

Table 5: **Parameter analysis of transition reasoning module:** We infer the initial motion snippet code of the current motion sequence from the previous motion and determine the optimal number of the inference code through experiments. The result shows that inferring the first two codes of current motion from the previous motion work best.

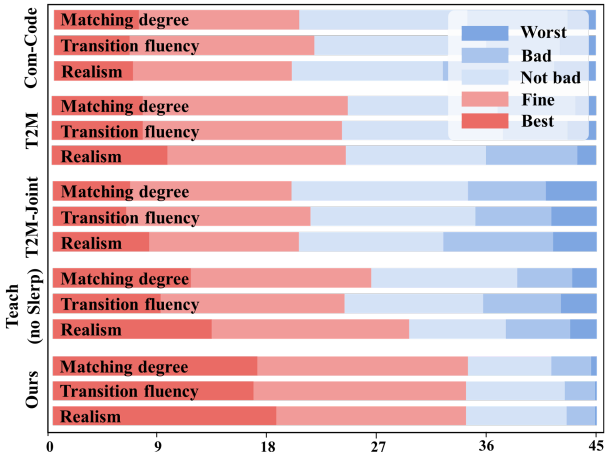

Figure 7: **User study:** We invite 45 users to score the results of our comparative experiments, and the color bars in the figure indicate the percentage of the scores. The results show that our method outperforms other baselines in terms of semantic matching, transition fluency, and realism.

current text.

The second experiment is to remove the L2G semantic fusion module. We only encode the features of the current text at each moment without considering the previous text. The experiment results show that the model's performances on R Precision, FID, and MultiModal Distance decrease significantly after removing the L2G semantic fusion module, indicating that previous text information is essential in synthesizing motion sequences that match context semantics. Although there is a slight improvement in the metric of transition fluency after the removal, the gap between them can not be observed from the naked eye. With the L2G semantic fusion module, our model shows better competence in evaluating the implicit relationship between current and previous text.

The third experiment is to remove the transition loss. The experimental results show that after removing the transition loss, the model's performances in all indicators have decreased, indicating that adding transition loss will guide the model to learn more semantic matching and transitional natural motion.

The fourth experiment does not use the segmented training strategy (explained in the supplementary file). The specific approach is that all motion sequences are extended to the maximum length (196 frames) with 0 in the training strategy. The experimental results show that when the segmented training strategy is not used, the performance of each indicator decreases, indicating that the segmented training strategy can make the model learn more thoroughly.

4.5. Parameter Analysis

In the transition reasoning module, we verify through experiments that we need to infer the number of motion snippet codes from the previous motions at the current initial stage. As shown in Tab. 5, the reasoning motion snippet code number of 1, 2, 3, and 5 are compared. The experiment results show that two motion snippet codes at the beginning of reasoning are optimal because the model can not fully capture the influence of previous motion information by reducing motion snippets inferred from previous motions. At

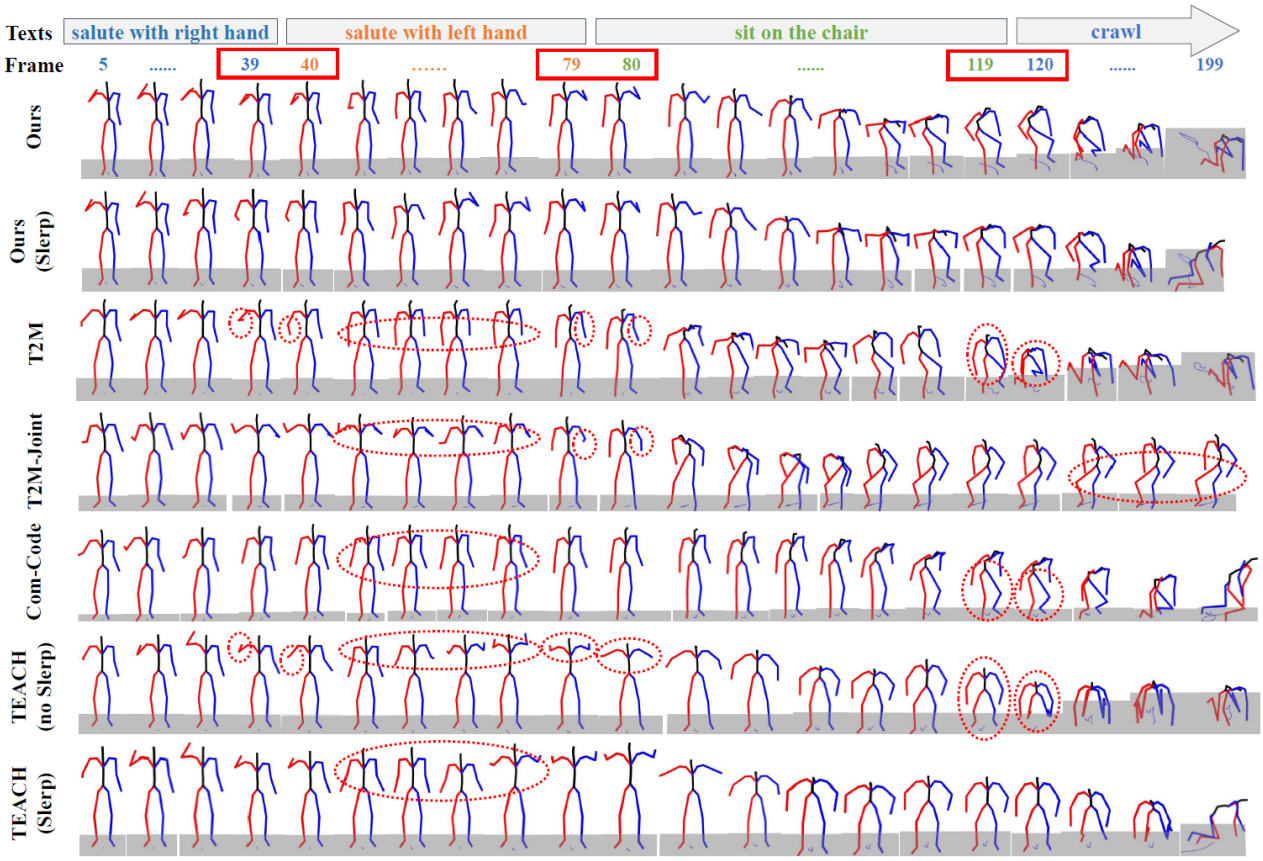

Figure 8: **Qualitative evaluation:** Continuously inputting four texts to show the visual results of our method and other baseline methods, except for the two frames between adjacent motions, the rest only show the keyframes. The results show that the motion synthesized by our method is the best in terms of semantic matching, transition fluency, and realism. More results are in the supplementary file.

the same time, the model will introduce too much previous motion information into the current motion if increasing the motion snippets inferred from previous motions.

4.6. Qualitative Evaluation

We make a qualitative comparison with the baselines, as shown in Fig. 8, our synthesized motion is the most natural in transition and nicely matches the semantics of the texts. Transition in T2M [11] is most evident in adjacent motion. T2M-Joint cannot fully express the content of the text description. When the input sequential texts contain more than two sentences, noticeable changes in the adjacent motion transition will occur. Complusion-Code produces a slight change in the transition. There is some semantic mismatch in TEACH [6] and abrupt transitions in the no Slerp version.

5. Conclusions and Future Works

We have advocated a novel framework for synthesizing semantic motion sequences driven by sequential texts while ensuring natural transitions between adjacent motions. Our approach leveraged a local-to-global semantic fusion module and a transition reasoning module to overcome the challenges of semantic association and transition fluency. Extensive experiments show that our method outperforms state-of-the-art techniques in terms of semantic matching, transition fluency, and realism.

However, our method still has limitations, such as difficulty in synthesizing head motions like "nodding" and "shaking head". In future work, we plan to improve this by exploring a way to assign different attention weights to different joints automatically. Additionally, since our STDM dataset is synthesized while not captured, we will further improve the data quality and expand the dataset scale.

Acknowledgements

This paper is supported by National Natural Science Foundation of China (62272021, 62102036), Beijing Natural Science Foundation (4222024), R&D Program of Beijing Municipal Education Commission (KM202211232003), Open Project Program of State Key Laboratory of Virtual Reality Technology and Systems, Beihang University (No.VRLAB2022A02), Research Unit of Virtual Human and Virtual Surgery (2019RU004).

References

[1] Mixamo. https://www.mixamo.com. Accessed February 17, 2023.

[2] CMU graphics lab motion capture database. http://mocap.cs.cmu.edu/, 2003.

[3] Chaitanya Ahuja, Dong Won Lee, Yukiko I.Nakano, et al. Style transfer for co-speech gesture animation: A multi-speaker conditional-mixture approach. European Conference on Computer Vision, pages 248–265, 2020.

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

[5] Simon Alexanderson, Gustav Eje Henter, Taras Kucherenko, et al. Style-controllable speech-driven gesture synthesis using normalising flows. Computer Graphics Forum, 39(2):487–496, 2020.

[6] Nikos Athanasiou, Mathis Petrovich, Michael J.Black, et al. TEACH: Temporal action composition for 3d humans. 2022 International Conference on 3D Vision (3DV), pages 414–423, 2022.

[7] Pablo Cervantes, Yusuke Sekikawa, Ikuro Sato, et al. Implicit neural representations for variable length human motion generation. European Conference on Computer Vision, pages 356–372, 2022.

[8] Jihoon Chung, Cheng hsin Wuu, Hsuan ru Yang, et al. HAA500: Human-centric atomic action dataset with curated videos. IEEE/CVF International Conference on Computer Vision, pages 13465–13474, 2021.

[9] Anindita Ghosh, Noshaba Cheema, Cennet Oguz, et al. Synthesis of compositional animations from textual descriptions. IEEE/CVF International Conference on Computer Vision, pages 1396–1406, 2021.

[10] Shiry Ginosar, Amir Bar, Gefen Kohavi, et al. Learning individual styles of conversational gesture. IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 3497–3506, 2019.

[11] Chuan Guo, Shihao Zou, Xinxin Zuo, et al. Generating diverse and natural 3d human motions from text. IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 5152–5161, 2022.

[12] Chuan Guo, Xinxin Zuo, Sen Wang, et al. Action2motion: Conditioned generation of 3d human motions. 28th ACM International Conference on Multimedia, pages 2021–2029, 2020.

[13] Chuan Guo, Xinxin Zuo, Sen Wang, et al. Action2video: Generating videos of human 3d actions. International Journal of Computer Vision, 130(2):285–315, 2022.

[14] Chuan Guo, Xinxin Zuo, Sen Wang, et al. TM2T: Stochastic and tokenized modeling for the reciprocal generation of 3d human motions and texts. European Conference on Computer Vision, pages 580–597, 2022.

[15] Ikhsanul Habibie, Weipeng Xu, Dushyant Mehta, et al. Learning speech-driven 3d conversational gestures from video. 21st ACM International Conference on Intelligent Virtual Agents, pages 101–108, 2021.

[16] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2:1735–1742, 2006.

[17] Yanli Ji, Feixiang Xu, Yang Yang, et al. A large-scale RGB-D database for arbitrary-view human action recognition. 26th ACM International Conference on Multimedia, pages 1510–1518, 2018.

[18] Jihoon Kim, Jiseob Kim, and Sungjoon Choi. Flame: Free-form language-based motion synthesis & editing. arXiv preprint arXiv:2209.00349, 2022.

[19] Taras Kucherenko, Patrik Jonell, Sanne van Waveren, et al. Gesticulator: A framework for semantically-aware speech-driven gesture generation. 2020 International Conference on Multimodal Interaction, pages 242–250, 2020.

[20] Yuanzhi Liang, Qianyu Feng, Linchao Zhu, et al. SEEG: Semantic energized co-speech gesture generation. IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 10473–10482, 2022.

[21] Xiao Lin and Mohamed R Amer. Human motion modeling using dvgans. arXiv preprint arXiv:1804.10652, 2018.

[22] Haiyang Liu, Zihao Zhu, Naoya Iwamoto, et al. BEAT: A large-scale semantic and emotional multi-modal dataset for conversational gestures synthesis. arXiv preprint arXiv:2203.05297, 2022.

[23] Jun Liu, Amir Shahroudy, Mauricio Perez, et al. NTU rgb+d 120: A large-scale benchmark for 3d human activity understanding. IEEE Transactions on Pattern Analysis and Machine Intelligence, 42(10):2684–2701, 2019.

[24] Xian Liu, Qianyi Wu, Hang Zhou, et al. Learning hierarchical cross-modal association for co-speech gesture generation. IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 10462–10472, 2022.

[25] Matthew Loper, Naureen Mahmood, Naureen Mahmood, et al. SMPL: A skinned multi-person linear model. Acm Transactions on Graphics, 34(6cd):248, 2015.

[26] Wei Mao, Miaomiao Liu, and Mathieu Salzmann. Weakly-supervised action transition learning for stochastic human motion prediction. IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 8151–8160, 2022.

[27] Dario Pavllo, David Grangier, and Michael Auli. Quaternet: A quaternion-based recurrent model for human motion. arXiv preprint arXiv:1805.06485, 2018.

[28] Mathis Petrovich, Michael J.Black, and Gül Varol. Action-conditioned 3d human motion synthesis with transformer vae. IEEE/CVF International Conference on Computer Vision, pages 10985–10995, 2021.

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

[30] Matthias Plappert, Christian Mandery, and Tamim Asfour. The KIT motion-language dataset. Big Data, 4(4):236–252, 2016.

[31] Alec Radford, Jong Wook Kim, Chris Hallacy, et al. Learning transferable visual models from natural language supervision. International Conference on Machine Learning, pages 8748–8763, 2021.

[32] Abhinanda R.Punnakkal, Arjun Chandrasekaran, Nikos Athanasiou, et al. BABEL: Bodies, action and behavior with english labels. IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 722–731, 2021.

[33] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, et al. Ntu rgb+ d: A large scale dataset for 3d human activity analysis. IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 1010–1019, 2016.

[34] Guy Tevet, Brian Gordon, Amir Hertz, et al. MotionCLIP: Exposing human motion generation to clip space. arXiv preprint arXiv:2203.08063, 2022.

[35] Sijie Yan, Zhizhong Li, Yuanjun Xiong, et al. Convolutional sequence generation for skeleton-based action synthesis. IEEE/CVF International Conference on Computer Vision, pages 4394–4402, 2019.

[36] Youngwoo Yoon, Bok Cha, Joo-Haeng Lee, et al. Speech gesture generation from the trimodal context of text, audio, and speaker identity. ACM Transactions on Graphics, 39(6):1–16, 2020.

[37] Mingyuan Zhang, Zhongang Cai, Liang Pan, et al. Motiondiffuse: Text-driven human motion generation with diffusion model. arXiv preprint arXiv:2208.15001, 2022.

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

[39] Hang Zhao, Antonio Torralba, Lorenzo Torresani, et al. Hacs: Human action clips and segments dataset for recognition and temporal localization. IEEE/CVF International Conference on Computer Vision, pages 8668–8678, 2019.

[40] Yi Zhou, Connelly Barnes, Jingwan Lu, et al. On the continuity of rotation representations in neural networks. IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 5745–5753, 2019.