在虚拟现实中实现非语言互动社交化身的新方法

艾力

2024 年 6 月

在虚拟现实中实现非语言互动社交化身的新方法

艾力

北京理工大学

中图分类号：TQ028.1
UDC 分类号：540

在虚拟现实中实现非语言互动社交化身的新方法

作 者 姓 名　艾力
学 院 名 称　计算机学院
指 导 教 师　老师
答辩委员会主席　老师
申请学位级别　工学硕士
学 科 专 业　计算机科学与技术
学位授予单位　北京理工大学
论文答辩日期　2024 年 6 月 1 日

Novel Methods for Realizing Non-Verbal Interactive Social Avatars in Virtual Reality

Candidate Name:	Manjotho Ali Asghar
School or Department:	Computer Science
Faculty Mentor:	Prof.
Chair, Thesis Committee:	Prof.
Degree Applied:	Doctor of Engineering
Major:	Computer Science and Technology
Degree by:	Beijing Institute of Technology
The Date of Defence:	June, 1st, 2024

研究成果声明

本人郑重声明：所提交的学位论文是我本人在指导教师的指导下进行的研究工作获得的研究成果。尽我所知，文中除特别标注和致谢的地方外，学位论文中不包含其他人已经发表或撰写过的研究成果，也不包含为获得北京理工大学或其它教育机构的学位或证书所使用过的材料。与我一同工作的合作者对此研究工作所做的任何贡献均已在学位论文中作了明确的说明并表示了谢意。

特此申明。

签　　名：　　日　期：

关于学位论文使用权的说明

本人完全了解北京理工大学有关保管、使用学位论文的规定，其中包括：① 学校有权保管、并向有关部门送交学位论文的原件与复印件；② 学校可以采用影印、缩印或其它复制手段复制并保存学位论文；③ 学校可允许学位论文被查阅或借阅；④ 学校可以学术交流为目的, 复制赠送和交换学位论文；⑤ 学校可以公布学位论文的全部或部分内容（保密学位论文在解密后遵守此规定）。

签　　名：　　日　期：

指导老师签名：　　日　期：

摘要

人类运动的理解在各个领域都至关重要，包括计算机图形学、机器人技术、医疗保健和交互式模拟。然而，传统方法通常难以有效地弥合动作语义和行为理解之间的差距，这是由于表示不足和运动层次结构中固有的不确定性所导致的。本论文提出了创新性的解决方案，以增强对人类运动的理解和生成。

首先，我们深入研究了人类运动理解的复杂性，并引入了一种称为模糊定性运动学（FQK）的新方法。这一开创性的框架将模糊推理和定性推理相结合，以便于对富有表现力和不确定性的运动事实进行建模。通过为各种运动属性创建模糊语言变量、术语和隶属函数，FQK 实现了从运动序列中提...

关键词：人类运动理解；社交化头像；模糊定性运动学；量化运动令牌。

Abstract

Understanding human motion is crucial in various fields, including computer graphics, robotics, healthcare, and interactive simulations. However, traditional methods often struggle to effectively bridge the gap between motion semantics and action comprehension due to inadequacies in representations and inherent uncertainties in kinematic hierarchies. This thesis proposes innovative solutions for enhancing the understanding and generation of human motion.

First, we delve into the complexities of human motion comprehension and introduce a novel approach termed Fuzzy Qualitative Kinematics (FQK). This groundbreaking framework combines fuzzy inference and qualitative ...

Keywords: Human motion understanding; social avatars; fuzzy qualitative kinematics; quantized motion tokens.

Table of Contents

Figures

Tables

Notations

BIT	北京理工大学的英文缩写
LaTeX	一个很棒的排版系统
LaTeX 2_ε	一个很棒的排版系统的最新稳定版
ctex	成套的中文 LaTeX 解决方案，由一帮天才们开发
$e^{\pi i} + 1 = 0$	一个集自然界五大常数一体的炫酷方程

Chapter 1 Introduction

1.1 Background

Human motion generation using skeleton-based 3D motion data has gained significant attention from researchers across various fields, including computer graphics, robotics, human-computer interaction (HCI), and computer vision. The primary objective of human motion generation is to develop models that are capable of synthesizing realistic, natural, and diverse human motions. These synthesized motions find applications across a broad spectrum, including digital film production, games, augmented/virtual reality (AR/VR), sports analysis, smart healthcare, human-robot interaction (HRI), and the creation of intelligent virtual avatars (IVAs).

Typically, a human motion generation task is conditioned on various factors, such as text[1], audio[1], scene[1], or another motion sequence[1]. In recent years, there has been a significant surge in the development of diverse generative methods, primarily driven by advancements in deep learning[1]. These approaches include Autoregressive models[1], Variational Autoencoders (VAE)[1], Normalizing Flows[1], Generative Adversarial Networks (GAN)[1], and Denoising Diffusion Probabilistic Models (DDPM)[1]. These methodologies have demonstrated their effectiveness in various domains such as text[1],[1], imagery[1], video[1], and 3D objects[1]. Additionally, significant progress in human motion modelling[1] has enabled the extraction of human motion from videos[1], as well as the creation of extensive human motion datasets[1]. Consequently, there has been growing interest within the research community in data-driven approaches for human motion generation in recent times.

Human motion generation presents several challenges ...

1.2 Motivation

Generating accurate 3D motions is a hot topic in the field of computer vision. Researchers are proposing various methods to generate human motions, such as predicting motions based on historical sequences, generating motions for a particular action class, and creating motions that follow a trajectory or a song. These methods aim to produce increasingly realistic and complex motions. There are numerous potential applications for these generative methods in

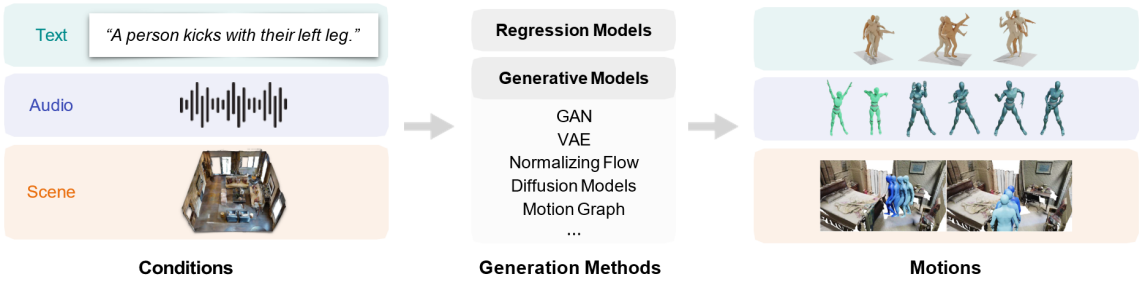

Figure 1.1 An overview of typical human motion generation approaches. Example images adapted from[1].

various areas, which has sparked the interest of researchers.

1.2.1 Human-Robot Interaction (HRI)

Safe and Efficient Collaboration: The utilization of human motion generation in training robots to predict human movements results in enhanced collaboration efficiency and safety within shared work environments. The simulation of various human actions enables robots to adapt their behavior and movements accordingly, thereby reducing the occurrence of collisions or accidents.

Enhanced Robot Configuration: Through the generation of diverse and intricate human motions, scholars can assess and enhance the design of robots. This process involves examining a robot's capacity to maneuver ...

1.3 Research Gaps in Existing Studies

Despite the progress made in human motion generation, there exist significant gaps and opportunities for further research and development. We identify the following limitations and lack of research in the domain:

Bridging action semantics and raw motion: The fundamental challenge lies in establishing a meaningful connection between the raw motion space and the action semantic space[1]. Conventionally, raw motion is represented as a sequence of 3D poses[1] or SMPL model parameters[1], while action semantics are characterized by action categories or textual descriptions in natural language. Existing approaches for human motion generation often struggle to

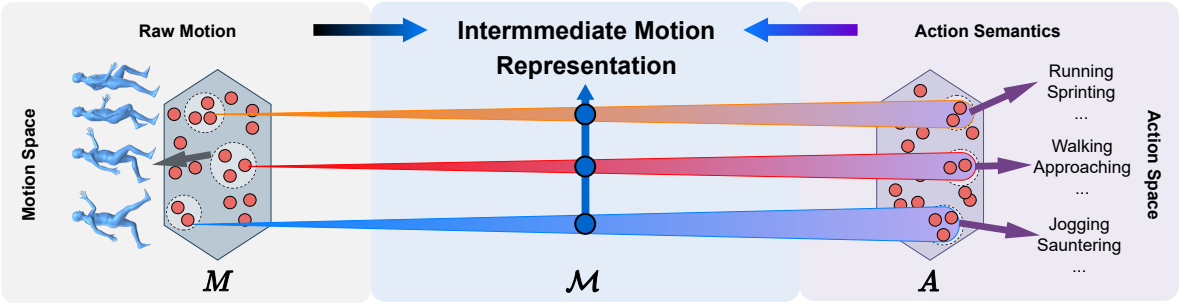

Figure 1.2 Visualizing semantic gap between action semantics and raw motion.

produce realistic motion sequences due to significant gap in mapping from action semantic text to raw motion pose sequence. Creating stronger connection between action semantics and raw motion with the aid of intermediate motion representations can make human motion understand more intuitive and improve the performance of underlying motion tasks.

Human reaction motion generation: Existing research mostly focus on generating human motion for a single character, while neglecting the motion sequences where interaction of two persons are involved. There is an opportunity for creating methods that generate reactive motion of one character when the action sequence of other is given.

Human interaction generation: Existing methods recognize and label the human interactions, but there is a lack of research pertaining to generation of motion sequences for interacting characters.

Accurate 3D human pose estimation: In human-agent interactions, 3D human pose estimation is fundamental. The quality of extracted 3D skeletons for human motion is directly proportional to the accuracy of motion models. Despite recent progress in 3D human pose estimation methods, practical applications often encounter challenges, particularly when dealing with complex poses commonly found in human motion sequences. Proposing an accurate 3DHPE method dealing with complex poses can benefit more accurate human motion generation models.

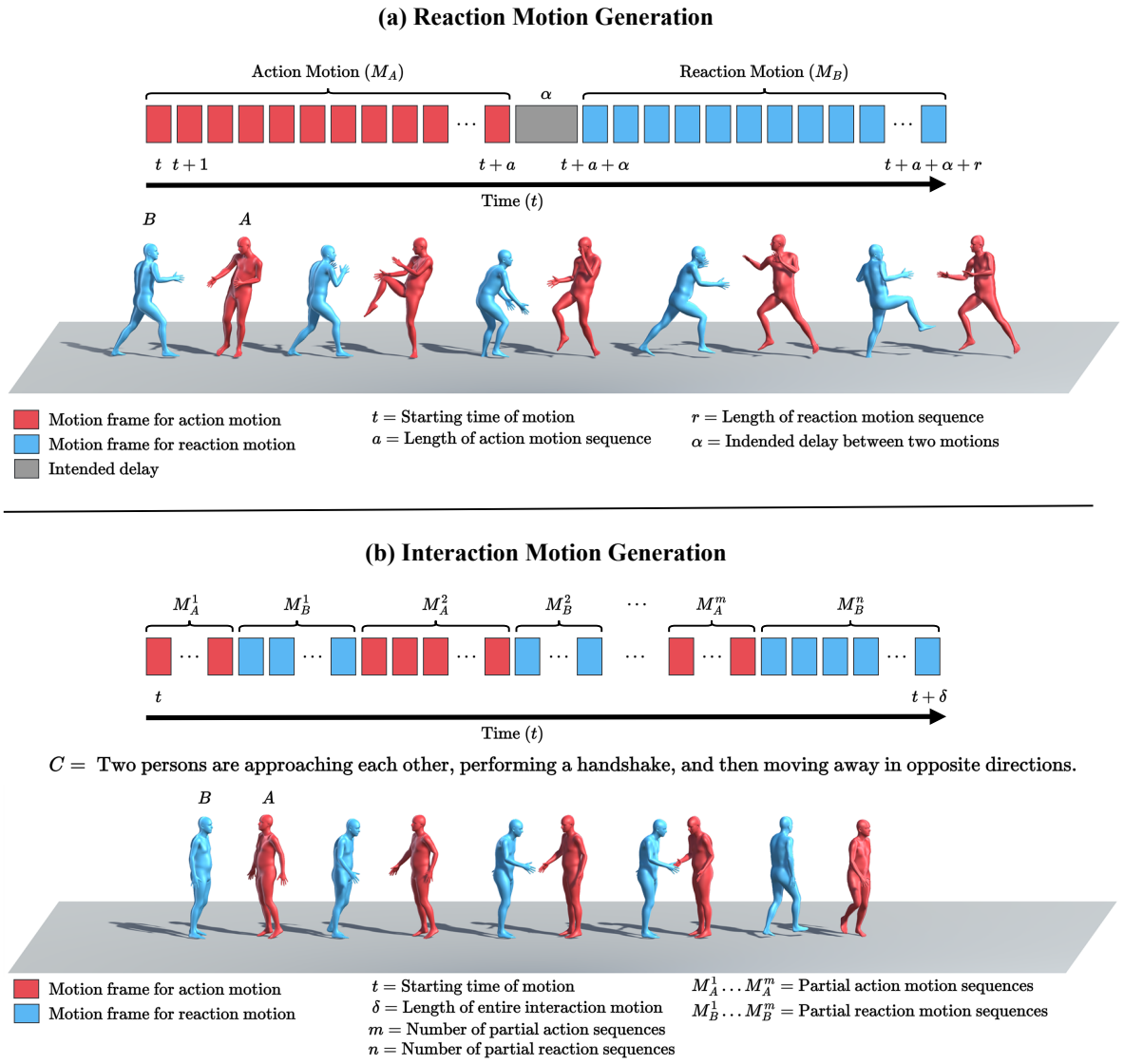

Figure 1.3 Illustration and mathematical symbols for various human motion generation process. The virtual character in red (Character A) represents the actor performing an action sequence. The character in blue (Character B) represents the actor performing a reaction sequence. (a) The illustration for reaction generation, when action motion sequence is given. (b) The illustration for interaction generation, given the conditioned signal C.

1.4 Problem Formulation

Consider the reaction motion generation concept illustrated in Fig.1.3 (a) and interaction motion generation shown in Fig. 1.3 (b), we define following problem statements for this thesis:

1. Developing an effective methodology to bridge action semantics A and raw motion M by generating an stronger intermediate motion representation $\mathcal{M}$. The intermediate representation is devised such that the mapping $A \mapsto \mathcal{M} \mapsto M$, ultimately leads to

motion generation characterized by improved precision, diversity, multi-modality, and accuracy across various motion generation tasks.

2. Given a action motion sequence M_B of virtual character B, develop a motion generator network to generate corresponding reaction motion sequence M_A of character A, where $M_A \in \mathbb{R}^{T\times J\times 3}$ and $M_B \in \mathbb{R}^{T\times J\times 3}$ are the sequences of skeleton poses with J number of joints and T number of frames.

3. Developing a motion generation network to simultaneously generate partial interactive motion sequences $M_A^1, \ldots, M_A^m$ and $M_B^1, \ldots, M_B^n$ for character A and character B, respectively, given a textual context C in a dyadic interaction scenario. Where each partial motion $M_A^M \in \mathbb{R}^{T\times J\times 3}$ and $M_B^N \in \mathbb{R}^{T\times J\times 3}$ are the sequences of skeleton poses with J number of joints and T number of frames (while T may be variable for each sequence).

4. Given a monocular RGB image I, develop an pose estimation network to output a 3D human pose P, where $P \in \mathbb{R}^{J\times 3}$ and J is the number of skeleton joints.

1.5 Research Questions

The study identifies and tries to answer the following research questions:

RQ1: How can intermediate motion representations effectively bridge the gap between action semantics and raw motion, improving the realism of generated motion sequences?

RQ2: What novel methods can be devised to generate reactive motion sequences for a character in response to the action sequence of another character?

RQ3: How can motion generation networks be designed and trained to accurately produce motion sequences for interacting characters in dyadic interactions?

RQ4: What advancements can be made in 3D human pose estimation to enhance the accuracy of extracted poses, particularly in dealing with complex poses commonly found in human motion sequences?

RQ5: How can a comprehensive framework integrating improved pose estimation, action-motion semantic comprehension, reaction motion modeling, and interactive motion modeling be designed and implemented to improve the naturalness and realism of generated human mo-

tion sequences?

1.6 Research Aim and Objectives

This aim of the study is to proposes a comprehensive framework to facilitate the generation of human motion sequences, incorporating coarse-to-fine pose estimation, action semantic comprehension, reaction motion modeling, and interactive motion modeling to improve the realism and naturalness of the generated motion sequences. The objective of the study are:

1. To identify and propose intermediate motion representation to effectively bridge action semantics and raw motion.
2. Designing and developing a network to generate the reaction motion sequence given the action motion sequence.
3. Designing and developing a network to generate interaction motion sequence in dyadic interactions.
4. Designing and developing an accurate 3D human pose estimation network to yield a 3D human pose given a monocular image.

1.7 Significance of the study

This research holds significant implications for numerous applications, including entertainment, education, healthcare, and human-computer ...

1.8 Contributions of the Study

This thesis makes several contributions to the field of human motion synthesis, including novel algorithms, methodologies, and insights that address key challenges and advance the state-of-the-art...

1.9 Thesis Organization

Chapter 2: This chapter provides a comprehensive ...

Chapter 3: This chapter introduces a novel methodology ...

Chapter 4: This chapter presents a novel approach ...

Chapter 5: This chapter employs anthropometric constraints ...

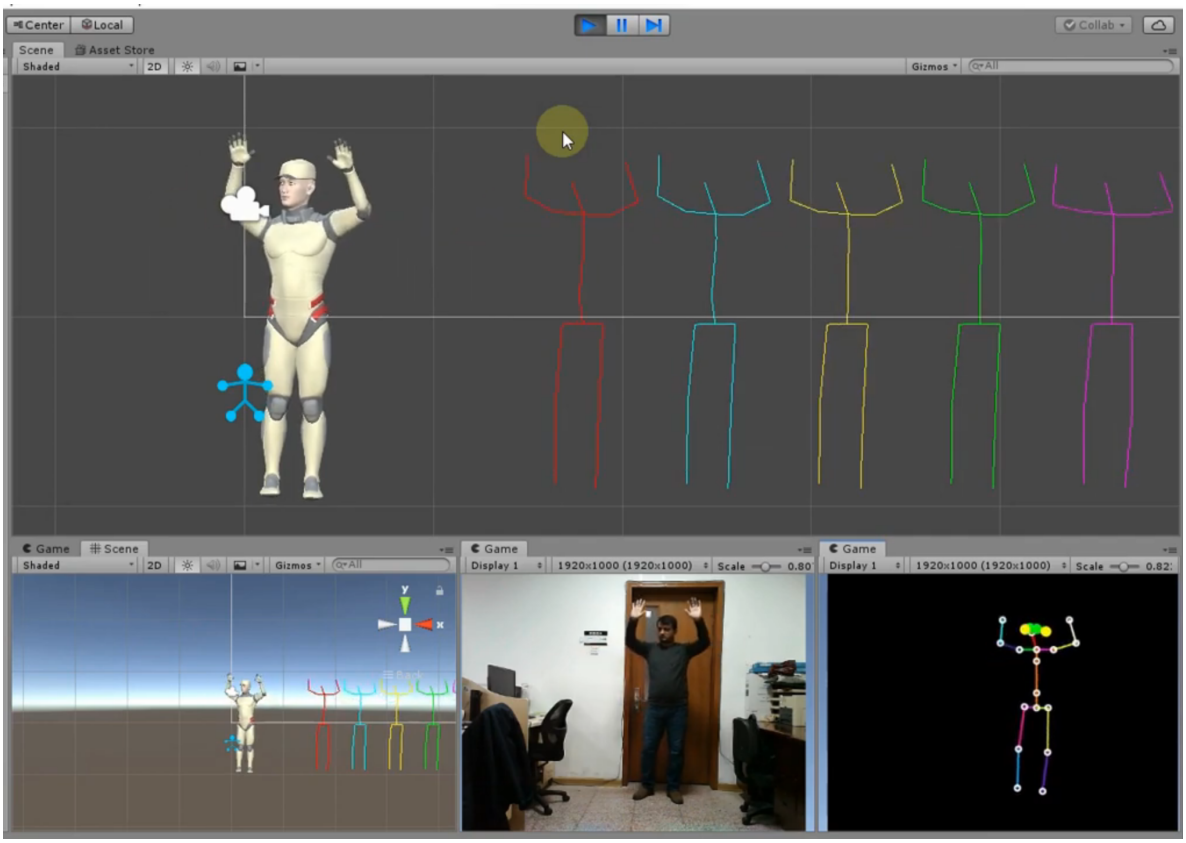

Figure 1.4 Human avatar interaction in virtual reality.

Chapter 2 Literature Review

Human beings exhibit a remarkable ability to plan and execute body movements in response to their intentions and environmental stimuli[1]. This intrinsic capability has become a central pursuit within artificial intelligence research, as there is a growing interest in the development of algorithms capable of generating motion patterns that mimic human behavior. This interdisciplinary endeavor has attracted attention from a myriad of research domains, such as computer vision[1], computer graphics[1], multimedia[1], robotics[1], and human-computer interaction[1]. The objective of human motion generation is to create natural, lifelike, and diverse human motions, which hold immense potential for application across various fields including film production, video games, augmented and virtual reality (AR/VR), human-robot interaction, and digital human representation.

2.1 Motion Data Representation

Human motion data is effectively represented by sequences of human body poses across the temporal ...

2.1.1 Keypoint-based Representation

In keypoint-based representation, the human body is depicted through a series of keypoints, denoting specific anatomical...

2.1.2 Rotation-based Representation

Another prevalent method for representing human pose involves joint angles, signifying the rotation of body parts or segments ...

2.1.3 The SMPL Model

The Skinned Multi-Person Linear (SMPL) model is parameterized by a set of pose and shape parameters, facilitating the generation of a 3D mesh representing a human body in a specific pose and shape (as illustrated in Figure **??**)...

Chapter 3 Connecting Action Semantics and Human Motion using Fuzzy Qualitative Kinematics

Human motion understanding is fundamental to various applications in computer graphics, human-robot interactions, digital environments, and entertainment. Existing approaches predominantly rely on modeling motion with quantitative or qualitative kinematic facts. However, they often struggle to establish a robust connection between motion and corresponding action semantics...

3.1 Introduction

Human motion understanding is integral to numerous applications, including human-robot interaction, automated computer animations, social humanoid in augmented/virtual reality, intelligent non-playing characters (NPCs) in video games, physical fitness, sport analysis, and digital film production[2-4]. The fundamental challenge lies in establishing a meaningful connection between the raw motion space and action semantic space ... .

3.2 Related Work

3.2.1 Human Motion Representation

An accurate representation of human motion is critical for generating realistic motion sequences[5]. Conventional motion is typically represented as a sequence of static pose representations characterized by joint...

3.3 Method

Effective mapping between the action and motion spaces is key to the success of any task involving motion. Fig. 3.1 visually demonstrates the semantic gap between these two ...

3.3.1 Quantized Fuzzy Membership Function

The standard fuzzy membership function assigns membership scores to crisp input values. However, the conventional real-value membership scale is computationally expensive. To address this issue, we introduce a quantized fuzzy membership function that strikes a balance between representational accuracy and computational complexity, as shown in Fig. **??**. This

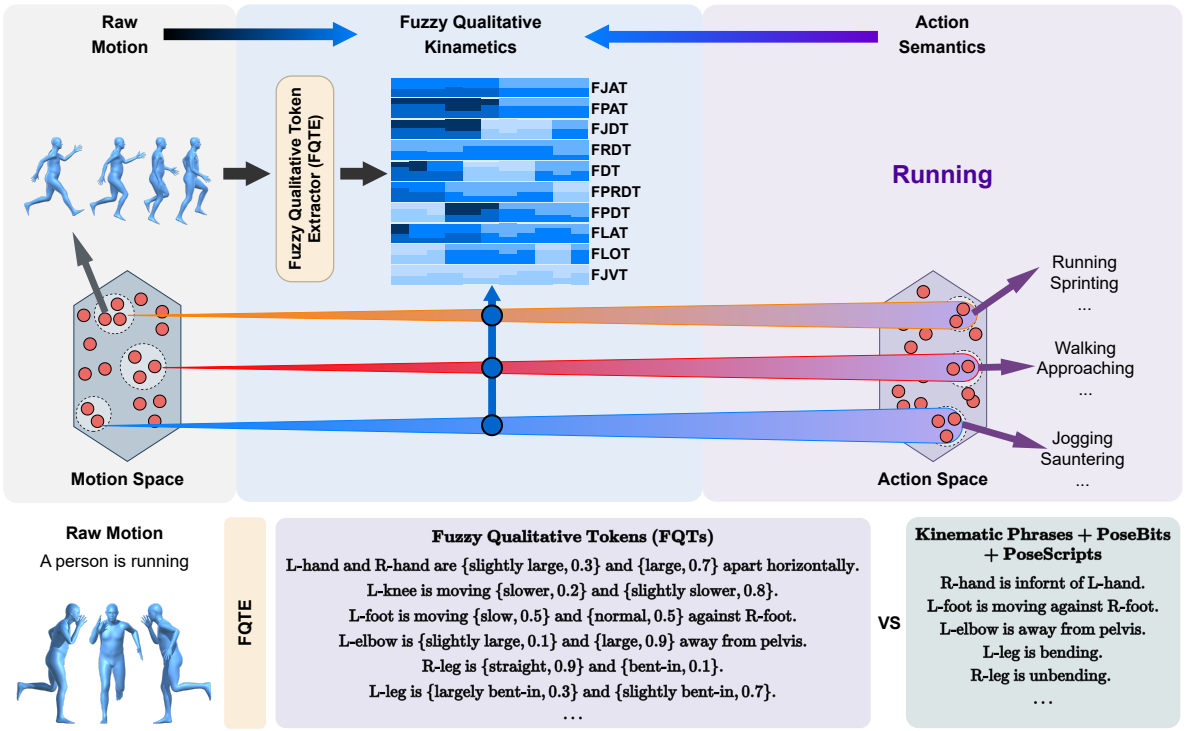

Figure 3.1 (Top) Illustrates the gap between two motion modalities, i.e., raw motion and action descriptions. Understanding human motion requires modeling a complex many-to-many mapping function between motion and action spaces. Fuzzy Qualitative Tokens (FQTs) are presented as an intermediate representation to bridge the gap (bottom) comparison of boolean kinematic facts used by previous studies[6-8] with our FQTs. FQTs provide expressive pose geometry and rich semantic information.

quantized fuzzy function extends the standard membership function by translating the real-valued membership scale into a quantized scale. The standard and quantized membership functions represented by the four-tuple $[a, b, c, d]$ are defined by eq (3.1) and (3.2), respectively.

$$\mu(x) = \begin{cases} 0 & \text{if } x \leq a \\ (x-a)/(b-a), & \text{if } a \leq x \leq b \\ 1, & \text{if } b \leq x \leq c \\ (d-x)/(d-c), & \text{if } c \leq x \leq d \\ 0, & \text{if } d \leq x \end{cases} \tag{3.1}$$

$$\dot{\mu}(x) = \lfloor \mu(x) \times \mu_{ql} \rfloor / \mu_{ql} \tag{3.2}$$

Where $\mu(.)$ and $\dot{\mu}(.)$ are the standard and quantized membership functions, respectively. The x denotes the crisp value, with μ_{ql} referring to the number of quantization levels.

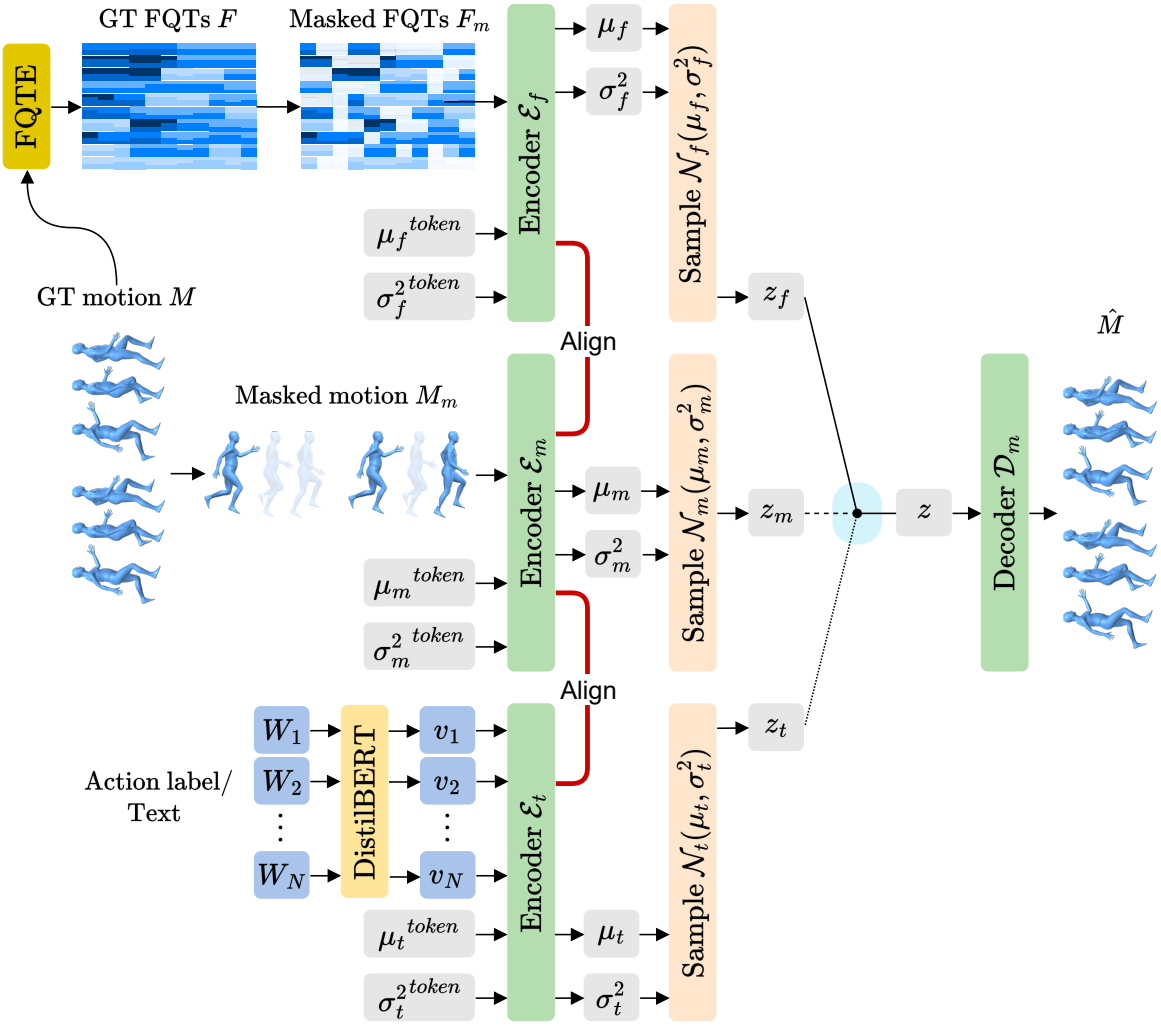

Figure 3.2 Method overview: During training, we encode FQTs, motion and text through their respective transformer encoders, together with modal-specific learnable distribution tokens. Each encoder outputs Gaussian distribution parameters, subject to KL losses, from which a latent vector z is sampled. The decoder uses the sampled variable to interpolate, predict, and generate a motion sequence.

Chapter 4 From Action to Reaction: Latent Space Regularization and Alignment for Human Reaction Motion Generation with Intermediate Motion Semantics

4.1 Summary

Creating lifelike virtual humanoids capable of simulating reactive movements towards humans or other characters holds ...

4.2 Introduction

Human motion generation is at the core of many applications in computer animation, augmented/virtual reality, robotics,...

4.3 Related Work

4.3.1 Human Motion Generation

The human motion generation task aims to produce realistic motion sequences for individuals or interactions involving humans...

4.4 Method

4.4.1 Problem Formulation

The human reaction-motion generation task aims to accurately generate the 3D skeleton-based motion of a character, given the action-motion sequence of another character. In the training phase, the input comprises action and reaction motion pairs $\left\{\left(x_a^{1:U}, x_r^{1:V}\right)\right\}$, where $x_a^{1:U}$ and $x_r^{1:V}$ are the action and corresponding reaction-motion sequences, respectively. The motion sequence $x^{1:T} = [S_1, \ldots, S_T]$ is represented as the human skeleton sequence defined for each discrete time interval $t \in \{1, \ldots, T\}$, where $S_t \in \mathbb{R}^{J\times 3}$ is the human skeleton configuration at any given time t with $J = 15$ 3D joint keypoints. Conversely, during the inference phase, our learned model can be expressed as a motion-motion translation function $G : x_a^{1:U} \rightarrow x_r^{1:V}$.

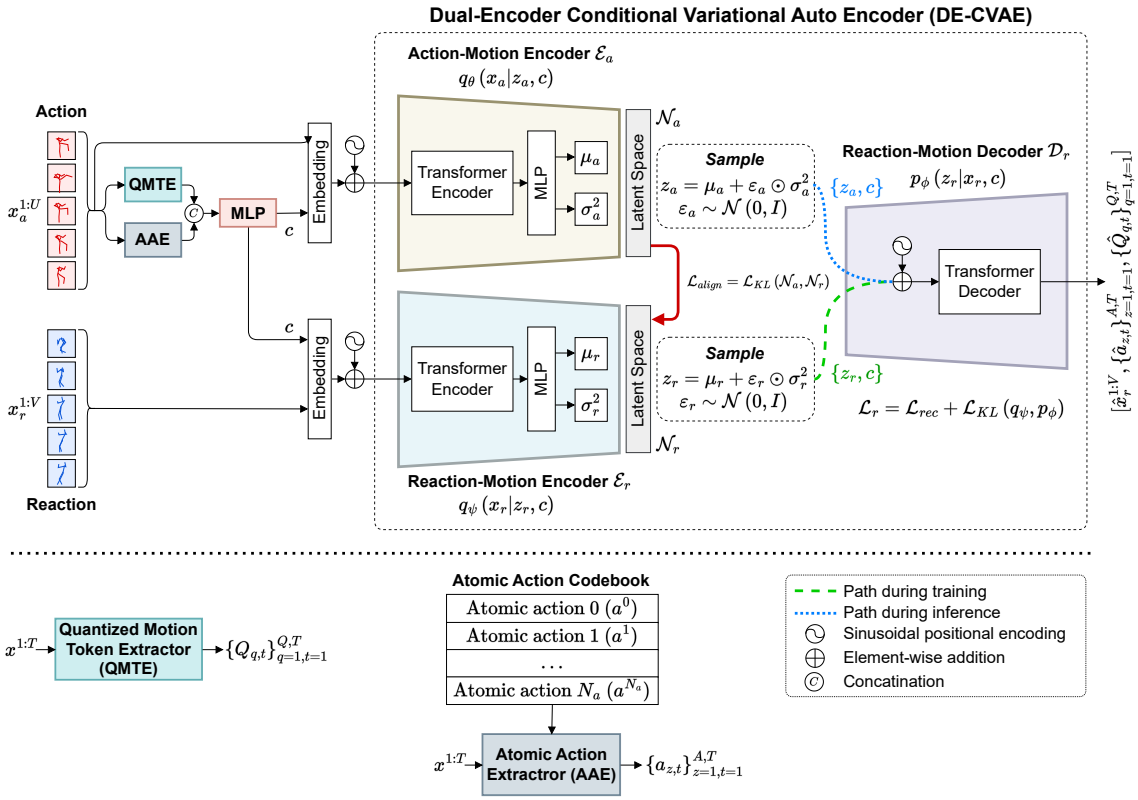

Figure 4.1 Overview of the proposed model (left) DE-CVAE network with two encoders and a decoder. QMTs and atomic action vectors are extracted from action-motion using QMTE and AAE modules, respectively (right) QMTE module, AAE module, and atomic action codebook.

4.4.2 Overview

The overall pipeline of the proposed method is shown in Fig. 4.1. The method uses an action-motion sequence $x_a^{1:U}$ and extracts quantized motion tokens and atomic action vectors to generate the conditional signal. The two encoders and a decoder use this conditional signal as a bias to disentangle and regularize the motion spaces. This results in a improved the action-reaction mapping. The reaction-motion encoder encodes ...

$$QPT_t^{(j,\dagger)} = \left\langle \frac{s_t^j - s_t^0}{q_l^p}, \frac{r_t^\dagger}{q_l^p} \right\rangle - \left\langle \frac{s_{t-1}^j - s_{t-1}^0}{q_l^p}, \frac{r_{t-1}^\dagger}{q_l^p} \right\rangle \tag{4.1}$$

4.5 Implementation Details

The model was implemented using the open-source PyTorch library. Network weights were initialized randomly and optimized using the Adam optimizer employing a mini-batch size

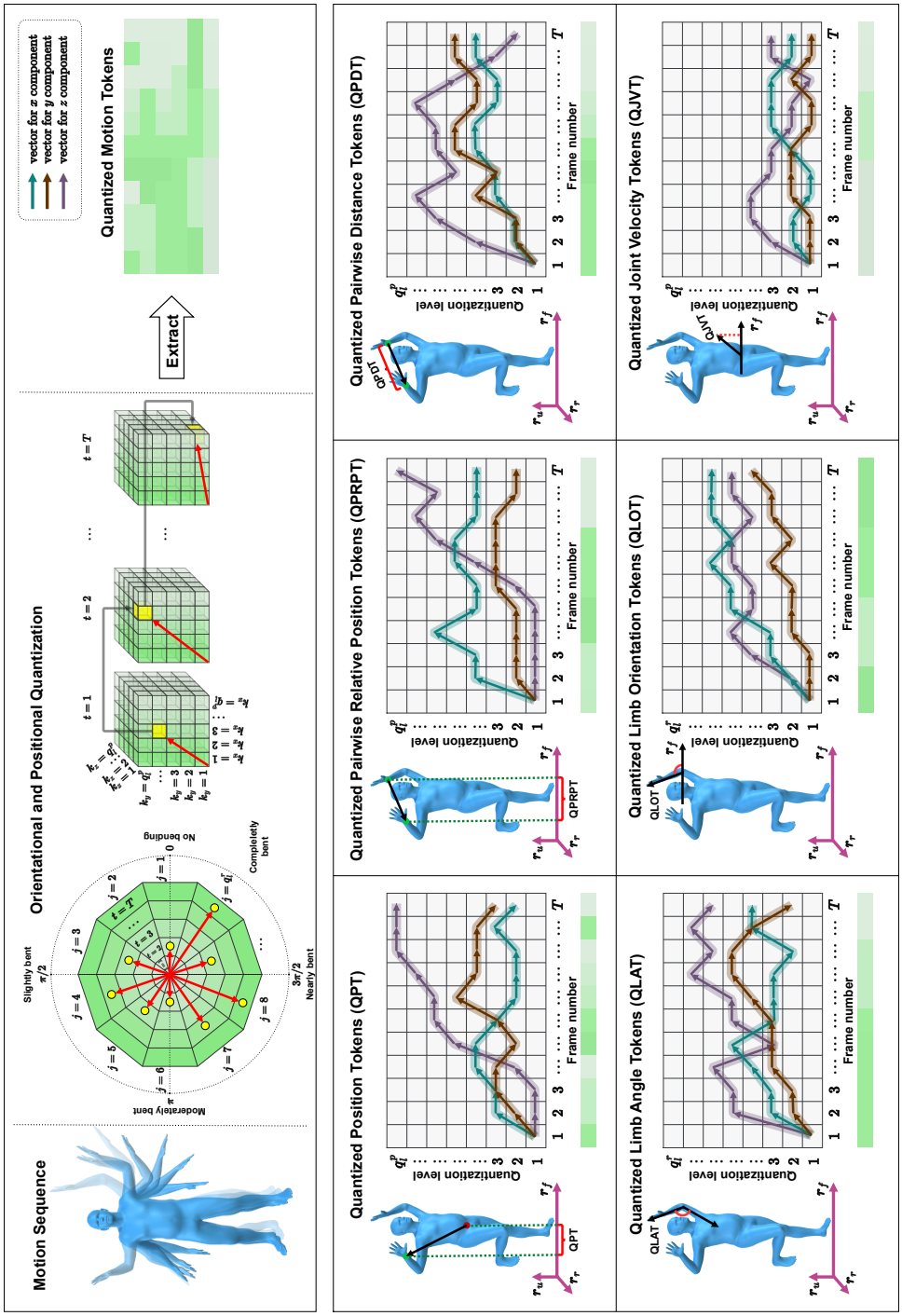

Figure 4.2 (Top left) Motion sequence (top middle) orientational and positional quantizations (top right) extracted quantized motion tokens (bottom) visual representations for QPT, QPRPT, QPDT, QLAT, QLOT, and QJVT.

of 32 for 45K epochs. The initial learning rate was set to $\alpha = 10^{-5}$ and decayed gradually with the Step LR scheduler with a step size of 8 and decay rate ...

4.6 Experiments

4.6.1 Datasets

SBU Kinect Interaction Dataset (SBU)[9] features a skeleton-based two-person interaction dataset collected from the actions of seven individuals organized in 21 pairs of two-actor sets...

Kinect-based 3D Human Interaction Dataset (K3HI)[10] offers a comprehensive repository of rich ...

4.6.2 Evaluation Metrics

Due to the inherent nature of reaction-motion generation, it is a one-to-many mapping problem...

The Fr'echet Motion Distance (FMD) is specifically designed to measure the distance between ground truth and generated data distribution...

4.6.3 State-of-the-Art Comparisons

Quantitative Evaluation. We conducted a comprehensive evaluation of the proposed method by comparing the FMD and diversity evaluation scores on the SBU, DuetDance, and K3HI datasets, as summarized in Table. **??**...

Qualitative Evaluation. We performed a qualitative evaluation by visually comparing the motion sequences generated by the proposed method with the groundtruth and various SOTA methods. Fig. 4.3, **??**, and **??** show the qualitative comparisons of the test samples from ...

Perceptual User Study. Relying solely on quantitative evaluation and self-performed qualitative analysis is insufficient due to the intricacies of the problem at hand. For a more comprehensive visual assessment of motion quality, ...

4.6.4 Ablation Study

To assess the effectiveness of the proposed modules in our network, we performed an ablation study on the SBU dataset, focusing on the diversity and FMD metrics. We set up four versions of our model and tested them with different ...

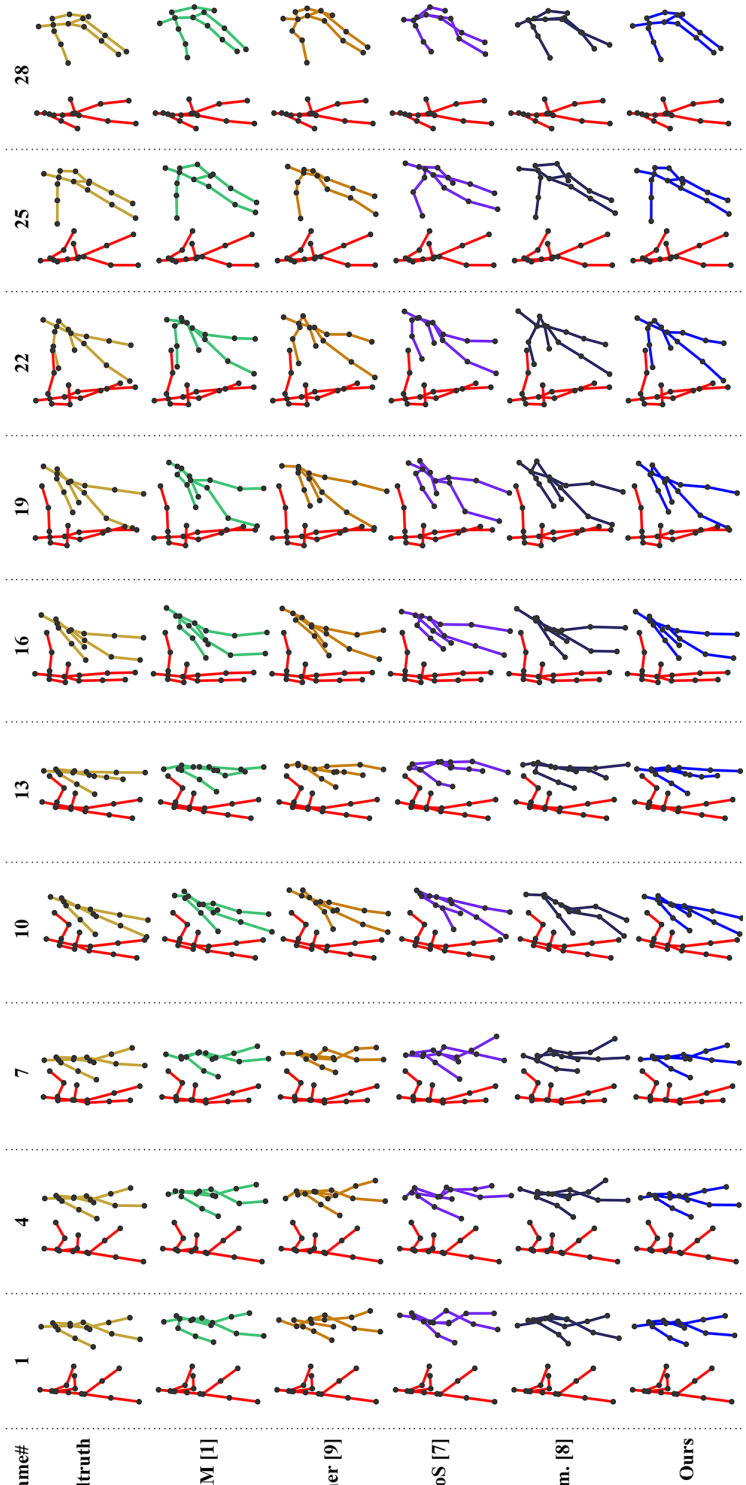

Figure 4.3 Visualization of motion generation on SBU dataset for punching class. Skeletons in red represent acting character, while the other colors correspond to the reacting character in various methods. (top to bottom) represent motion sequences for groundtruth, generated by methods[11],[12],[13],[14], and our results, respectively. (left to right) selective frames during temporal transition.

Table 4.1 (Left) Classification accuracy for each class in the SBU, Duetance, and K3HI datasets, comparing our method with groundtruth and exiting approaches ([11],[13],[14],[12],intergen, and[15]) (right) user perception study, comparing our methods with groundtruth and existing approaches ([12] and[11]) across the same datasets. " ↑ ": indicates higher is better. **Bold** specifies the best results.

Method	GT	GAN-RM [11]	ReMoS [13]	InterHum. [14]	InterFor. [12]	InterGen intergen	Social Rob. [15]	Ours	GT	InterFor. [12]	GAN-RM [11]	Ours
	Classification Accuracy (↑)								User Preference (↑)			
SBU Kinect Interact												
Walking	89	75	63	60	76	53	44	**85**	95%	77%	80%	**89%**
Kicking	100	**94**	79	68	82	75	77	93	92%	**85%**	78%	81%
Pushing	94	75	71	70	83	68	71	**91**	88%	77%	**78%**	73%
Shaking Hands	97	**91**	72	67	82	73	71	87	86%	70%	73%	**82%**
Exchanging	95	**89**	69	82	72	66	67	88	89%	78%	**88%**	76%
Punching	100	90	78	75	88	78	76	**93**	93%	72%	75%	**90%**
Average	95.83	85.67	72	70.33	80.5	68.83	67.67	**89.5**	90.5%	76.5%	78.67%	**81.83%**
DuetDance												
Cha-Cha	94	80	73	68	74	69	53	**87**	88%	74%	76%	**82%**
Jive	95	86	70	71	81	69	73	**93**	79%	62%	**76%**	74%
Rumba	92	78	66	67	78	67	64	**85**	78%	68%	66%	**74%**
Salsa	97	**93**	75	73	78	70	71	92	84%	66%	75%	**79%**
Samba	96	77	74	69	**89**	71	60	82	85%	75%	70%	**83%**
Average	94.8	82.8	71.6	69.6	80	69.2	64.2	**87.8**	82.8%	69%	72.6%	**78.4%**
K3HI												
Approaching	98	84	72	78	78	72	65	**97**	88%	74%	78%	**81%**
Departing	93	90	68	82	66	69	68	**93**	84%	72%	76%	**80%**
Kicking	100	82	72	93	76	75	70	**94**	89%	**86%**	76%	85%
Pushing	100	**96**	77	84	76	72	74	95	89%	78%	74%	**85%**
Shaking	96	**94**	68	75	75	71	71	90	87%	76%	78%	**84%**
Exchanging	93	77	72	84	70	69	48	**93**	90%	78%	**88%**	86%
Punching	100	89	75	90	82	71	74	**94**	90%	72%	80%	**89%**
Pointing	94	81	68	87	66	70	67	**89**	86%	73%	**86%**	84%
Average	96.75	86.62	71.5	84.12	73.62	71.125	67.12	**93.12**	87.87%	76.12%	79.5%	**84.25%**

4.7 Conclusion

In this paper, we introduce a novel DE-CVAE network featuring two encoders designed for the task of human reaction motion generation...

4.8 Future Work

Despite significant improvements in the generation task, our approach faces three major challenges. (1) As the interactions...

Conclusions

In this study, we address the long-standing challenge of bridging raw human motion and action semantics via fuzzy qualitative kinematics. By combining fuzzy logic and qualitative kinematic reasoning, we proposed a novel ...

References

[1] Gao X, Yang Y, Xie Z, et al. GUESS: GradUally Enriching SyntheSis for Text-Driven Human Motion Generation[J]. IEEE Transactions on Visualization and Computer Graphics, 2024.

[2] Wang Y, Leng Z, Li F W B, et al. Fg-T2M: Fine-Grained Text-Driven Human Motion Generation via Diffusion Model[C]. IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023. IEEE, 2023: 21978-21987.

[3] Liu Y, Zhang H, Li Y, et al. Skeleton-based Human Action Recognition via Large-kernel Attention Graph Convolutional Network[J]. IEEE Transactions on Visualization and Computer Graphics, 2023, 29(5): 2575-2585.

[4] Moullec Y, Cogné M, Saint-Aubert J, et al. Assisted walking-in-place: Introducing assisted motion to walking-by-cycling in embodied virtual reality[J]. IEEE Transactions on Visualization and Computer Graphics, 2023, 29(5): 2796-2805.

[5] Loi I, Zacharaki E I, Moustakas K. Machine learning approaches for 3D motion synthesis and musculoskeletal dynamics estimation: A Survey[J]. IEEE Transactions on Visualization and Computer Graphics, 2023.

[6] Liu X, Li Y, Zeng A, et al. Bridging the Gap between Human Motion and Action Semantics via Kinematic Phrases[J]. CoRR, 2023, abs/2310.04189.

[7] Pons-Moll G, Fleet D J, Rosenhahn B. Posebits for Monocular Human Pose Estimation[C]. 2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014. IEEE Computer Society, 2014: 2345-2352.

[8] Delmas G, Weinzaepfel P, Lucas T, et al. PoseScript: 3D Human Poses from Natural Language[C]. Lecture Notes in Computer Science: Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part VI: vol. 13666. Springer, 2022: 346-362.

[9] Yun K, Honorio J, Chattopadhyay D, et al. Two-person interaction detection using body-pose features and multiple instance learning[C]. 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, Providence, RI, USA, June 16-21, 2012. IEEE Computer Society, 2012: 28-35.

[10] Hu T, Zhu X, Guo W. Two-person interaction recognition based on key poses[J]. Journal of Computational Information Systems, 2014, 10: 1965-1972.

[11] Men Q, Shum H P H, Ho E S L, et al. GAN-based reactive motion synthesis with class-aware discriminators for human-human interaction[J]. Comput. Graph., 2022, 102: 634-645.

[12] Chopin B, Tang H, Otberdout N, et al. Interaction Transformer for Human Reaction Generation[J].

IEEE Transactions on Multimedia, 2023, 25: 8842-8854.

[13] Ghosh A, Dabral R, Golyanik V, et al. ReMoS: Reactive 3D Motion Synthesis for Two-Person Interactions[J]. CoRR, 2023, abs/2311.17057.

[14] Liu Y, Chen C, Yi L. Interactive Humanoid: Online Full-Body Motion Reaction Synthesis with Social Affordance Canonicalization and Forecasting[J]. arXiv preprint arXiv:2312.08983, 2023.

[15] Ko W, Jang M, Lee J, et al. Nonverbal Social Behavior Generation for Social Robots Using End-to-End Learning[J]. CoRR, 2022, abs/2211.00930.

[16] Manjotho A A, Tewolde T T, Niu Z. Connecting Action Semantics and Human Motion using Fuzzy Qualitative Kinematics[J]. IEEE Transactions on Fuzzy Systems, 2024. (SCI, Under Review).

[17] Manjotho A A, Tewolde T T, Niu Z. From Action to Reaction: Latent Space Regularization and Alignment for Human Reaction Motion Generation with Intermediate Motion Semantics[J]. IEEE Transactions on Multimedia, 2024. (SCI, Under Review).

[18] Manjotho A A, Tewolde T T, Niu Z. Coarse-to-Fine 3D Human Pose Estimation using Anthropometric Constraints and Synthetic Localization Errors[J]. IEEE Transactions on Intelligent Vehicles, 2024. (SCI, Under Review).

[19] Wang Q, Zhang K, Manjotho A A. Skeleton-Based ST-GCN for Human Action Recognition With Extended Skeleton Graph and Partitioning Strategy[J]. IEEE Access, 2022, 10: 41403-41410. (SCI, Published).

[20] Manjotho A A, Quanyu W, Kaixiang Z, et al. Pose correction using real-time pose buffer in a Human Pose Estimation (HPE) system (人体姿势估计系统中使用实时姿势缓冲区校正的方法)[M]. State Intellectual Property Office of the People's Republic of China, 2022. (CN115083018A, Accepted).

[21] Manjotho A A, Quanyu W, Yue S, et al. Pose stabilization in real-time 2D Human Pose Estimation (HPE) System (实时 2D 人体姿势估计系统中的姿势稳定系统和方法)[M]. State Intellectual Property Office of the People's Republic of China, 2022. (CN113762129A, Accepted).

[22] Manjotho A A, Quanyu W, Yue S, et al. Anthropometric limb depth corrector for 2D to 3D pose mapping in a 2D Human Pose Estimation (HPE) system (2D 人体姿态估计系统中基于人体测量的肢体校正器)[M]. State Intellectual Property Office of the People's Republic of China, 2022. (CN113643362A, Accepted).

[23] Manjotho A A, Quanyu W, Zhi W, et al. Head pose estimation via minimalistic facial landmark set in a real-time 2D Human Pose Estimation (HPE) System (一种通过简单面部关键点实时估计 2D 头部姿势的方法)[M]. State Intellectual Property Office of the People's Republic of China, 2022. (CN113837098A, Accepted).

Appendix A SMPL Parameter Settings

This is optional ...

Publications During Studies

一、Articles

[1] Wang Q, Zhang K, **Manjotho A A**. Skeleton-Based ST-GCN for Human Action Recognition With Extended Skeleton Graph and Partitioning Strategy[J]. IEEE Access, 2022, 10: 41403-41410. (SCI, Published).

[2] **Manjotho A A**, Tewolde T T, Niu Z. Coarse-to-Fine 3D Human Pose Estimation using Anthropometric Constraints and Synthetic Localization Errors[J]. IEEE Transactions on Intelligent Vehicles, 2024. (SCI, Under Review).

[3] **Manjotho A A**, Tewolde T T, Niu Z. Connecting Action Semantics and Human Motion using Fuzzy Qualitative Kinematics[J]. IEEE Transactions on Fuzzy Systems, 2024. (SCI, Under Review).

[4] **Manjotho A A**, Tewolde T T, Niu Z. From Action to Reaction: Latent Space Regularization and Alignment for Human Reaction Motion Generation with Intermediate Motion Semantics[J]. IEEE Transactions on Multimedia, 2024. (SCI, Under Review).

二、Patents

[1] **Manjotho A A**, et al. Pose correction using real-time pose buffer in a Human Pose Estimation (HPE) system (人体姿势估计系统中使用实时姿势缓冲区校正的方法)[M]. State Intellectual Property Office of the People's Republic of China, 2022. (CN115083018A, Accepted).

[2] **Manjotho A A**, et al. Anthropometric limb depth corrector for 2D to 3D pose mapping in a 2D Human Pose Estimation (HPE) system (2D 人体姿态估计系统中基于人体测量的肢体校正器)[M]. State Intellectual Property Office of the People's Republic of China, 2022. (CN113643362A, Accepted).

[3] **Manjotho A A**, et al. Pose stabilization in real-time 2D Human Pose Estimation (HPE) System (实时 2D 人体姿势估计系统中的姿势稳定系统和方法)[M]. State Intellectual Property Office of the People's Republic of China, 2022. (CN113762129A, Accepted).

[4] **Manjotho A A**, et al. Head pose estimation via minimalistic facial landmark set in a real-time 2D Human Pose Estimation (HPE) System (一种通过简单面部关键点实时估计 2D 头部姿势的方法)[M]. State Intellectual Property Office of the People's Republic of China, 2022. (CN113837098A, Accepted).

Acknowledgement

First and foremost, ...

Author Biography

The author ...