UnifiedGesture: A Unified Gesture Synthesis Model for Multiple Skeletons

Sicheng Yang,1, Zilin Wang*,1, Zhiyong Wu[1,4], Minglei Li[2], Zhensong Zhang[3], Qiaochu Huang [1], Lei Hao[3], Songcen Xu[3], Xiaofei Wu[3], Changpeng Yang[2], Zonghong Dai[2]*

[1] Tsinghua Shenzhen International Graduate School, Tsinghua University, China [2] Huawei Cloud Computing Technologies Co., Ltd, China [3] Huawei Noah's Ark Lab, China [4] The Chinese University of Hong Kong, Hong Kong SAR, China

1. Introduction

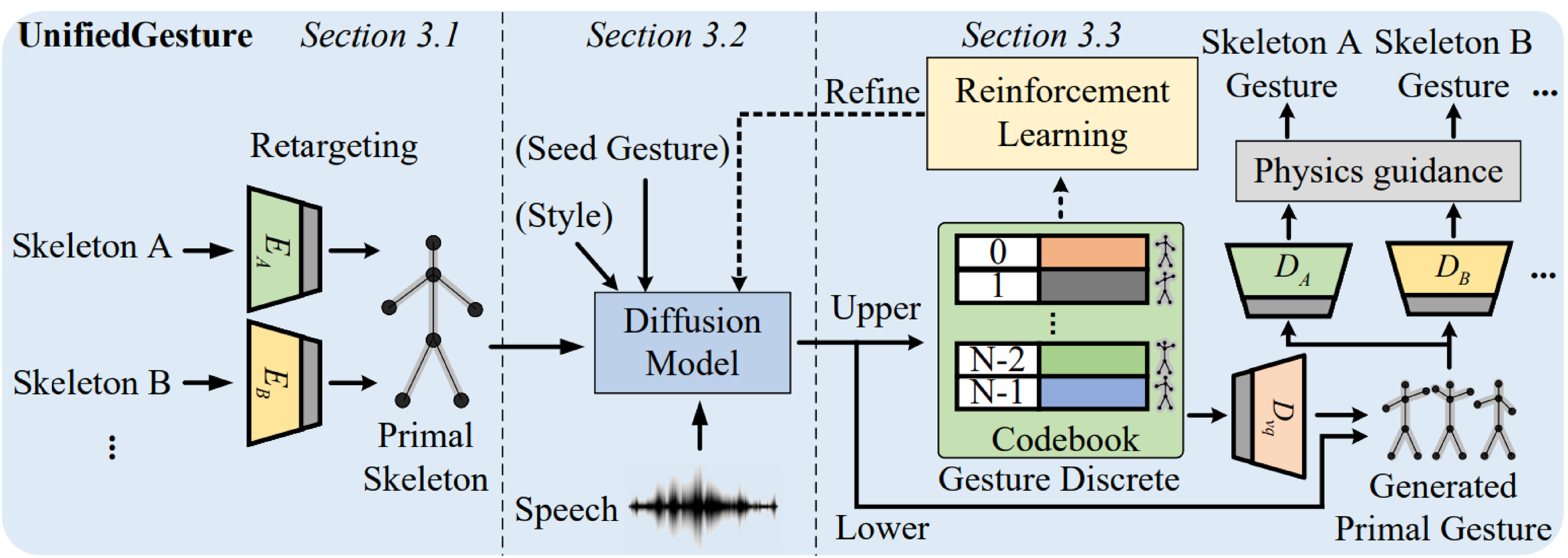
1.1 Motivation

- Goal:
  - ✓ Develop a comprehensive gesture synthesis model that can cater to multiple skeletons, ensuring natural and appropriate gestures in sync with speech
- Problem:
  - Existing gesture synthesis models are limited in their adaptability to various skeletons
  - The need for a model that can generate gestures that are both semantically relevant and natural in appearance

1.2 Contribution

- ✓ Introduction of a unified model that bridges the gap between different skeletal structures
- ✓ Present a temporally aware attention-based diffusion model on the primal skeleton for co-speech gesture generation
- ✓ Incorporation of advanced techniques like reinforcement learning and VQVAE to enhance gesture quality
- ✓ Extensive experiments show that our model can generate human-like, speech-matched, stylized, diverse, controllable, and physically plausible gestures

2. Methodology

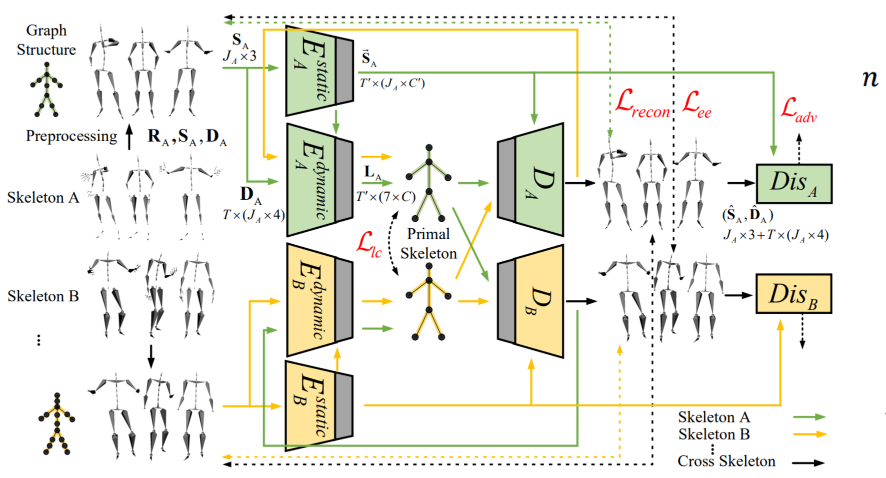


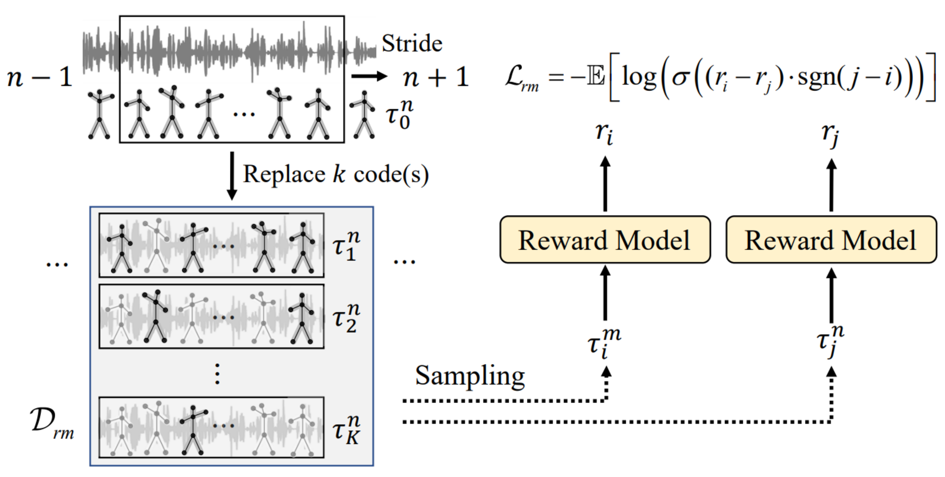
2.1 Multiple Skeletons Retargeting Network

The adjacency lists are expressed as $\mathcal{N}^d = \{\mathcal{N}_1^d, \mathcal{N}_2^d, \ldots, \mathcal{N}_J^d\}$

Two reference poses $\mathbf{P}_A$ and $\mathbf{P}_B$ can be aligned through global and local translation and rotation: $\mathbf{P}_B = \mathbf{Q}^{AB}\mathbf{P}_A\left(\mathbf{Q}^{AB}\right)^\top$

The motion of different skeletons consists of a static component $\mathbf{S} \in \mathbb{R}^{J\times 3}$ (joint offsets) and a dynamic one $\mathbf{D} \in \mathbb{R}^{T\times(J\times 4)}$ (joint rotations).To unify the motion of the different skeletons, we utilize a retargeting network architecture similar to [1]





2.2 Diffusion Model for Speech-driven Gesture Generation

Our goal is to synthesize a gesture $\mathbf{L}^{1:N}$ of length N given noising step t_d, noisy gesture $\mathbf{L}_{t_d}$ and conditions c (including audio a, style s, and seed gesture d). That is $\mathbf{L}_0 = \text{Denoise}\left(\mathbf{L}_{t_d}, t_d, c\right)$.

The Denoising module can be trained by optimizing the Huber loss between the generated poses $\mathbf{L}_0$ and the ground truth human gestures $\mathbf{L}_0$ on the training examples:

$$\mathcal{L}_{diff} = \lambda_{diff} E_{\mathbf{L}_0\sim q(\mathbf{L}_0|c), t_d\sim[1,T_d]}\left[\text{HuberLoss}(\mathbf{L}_0 - \mathbf{L}_0)\right]$$

2.3 Gesture Generation Refinement

2.3.1 Primal Gesture VQVAE

Each code represents a unique gesture. Besides, discrete spaces are more conducive to reinforcement learning for exploration. The VQVAE can be trained by optimizing $\mathcal{L}_{vq}$:

$$\mathcal{L}_{vq} = \left\|\mathbf{L}_0^{\text{upper}} - \mathbf{L}_0^{\text{upper}}\right\|_1 + \alpha_1\left\|\mathbf{L}_0^{\text{upper}'} - \mathbf{L}_0^{\text{upper}'}\right\|_1 + \alpha_2\left\|\mathbf{L}_0^{\text{upper}''} - \mathbf{L}_0^{\text{upper}''}\right\|_1 + \left\|\text{sg}[\mathbf{u}] - \mathbf{u_q}\right\| + \beta_{vq}\left\|\mathbf{u} - \text{sg}\left[\mathbf{u_q}\right]\right\|$$

2.3.2 Reinforcement Learning Finetuning

In this paper, we adopted Inverse Reinforcement Learning (IRL) to learn a neural network model from human demonstrations. Given the reward model, we use the REINFORCE algorithm to improve the model: $\mathcal{L}_{RL} = -\mathbb{E}_{\tau\sim\pi}\left[\log p_\pi(\tau) r(\tau)\right]$.

2.3.3 Physics Guidance

We consider that the foot should have contact with the ground when there is a left-right acceleration or an upward acceleration of the root. And we use standard Inverse Kinematics (IK) optimization for physics guidance.

3. Experiments

3.1 Experiment Preparation

- Retargeting network
  - Evaluation on the Trinity and ZEGGS datasets. d_{re}=4, then the primal gesture is 7.5 fps.
  - Adam optimizer with a batch size of 256 for 16000 epochs.
- Diffusion model
  - Gesture data are cropped to a length of N = 30 (4 seconds).
  - AdamW optimizer (learning rate is 3×10^{-5}) with a batch size of 256 for 1000000 steps.
- VQVAE
  - The size C_b of codebook $\mathcal{Z}_u$ is set to 512 with dimension n_z is 512. Down-sampling rate d_{vq}=2.
  - ADAM optimizer (learning rate is e^{-4}, β_1 = 0.5, β_2 = 0.98) with a batch size of 128 for 200 epochs.

3.2 Comparison to Existing Methods

- *Human-likeness*. Our model significantly surpasses the compared state-of-the-art methods.However, it is not significantly different from ExampleGestures.
- *Gesture and speech appropriateness*. Our model significantly outperforms StyleGestures, Audio2Gesture, and ExampleGestures, giving competitive results with DiffuseStyleGesture.

Name	Objective evaluation				Subjective evaluation	
	Global CCA	CCA for each sequence	FGD ↓	Diversity ↑	Human-likeness	Appropriateness
Ground Truth	1.000	1.00 ± 0.00	0.0	10.03	4.22 ± 0.11	4.22 ± 0.11
StyleGestures [4]	0.978	**0.98 ± 0.01**	15.89	**13.86**	3.56 ± 0.12	3.17 ± 0.13
Audio2Gesture [43]	0.969	0.97 ± 0.01	19.78	6.148	3.61 ± 0.11	3.15 ± 0.14
ExampleGestures [19]	0.914	**0.98 ± 0.01**	10.49	5.418	3.77 ± 0.12	3.17 ± 0.14
DiffuseStyleGesture [85]	0.987	0.97 ± 0.01	11.98	11.22	3.66 ± 0.12	**3.46 ± 0.14**
Ours	**0.988**	0.95 ± 0.02	**3.850**	7.039	**3.80 ± 0.11**	3.42 ± 0.14

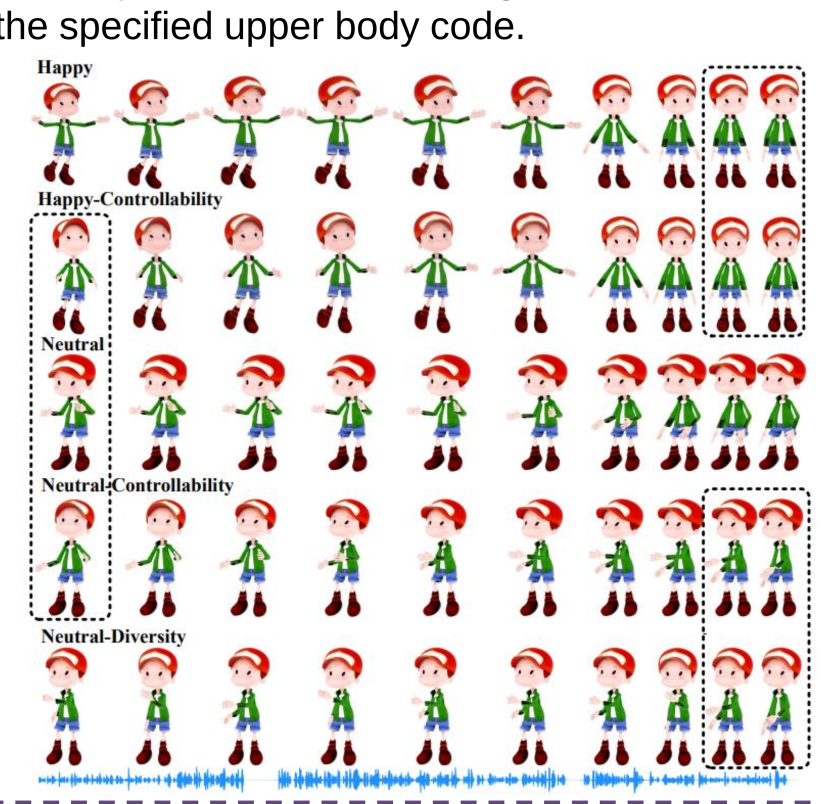
3.3 Ablation Studies

- *Human-likeness*. the scale of the dataset has a significant effect on the results, which shows the importance of unifying the gesture dataset.
- *Speech and gesture appropriateness*. The scale of the dataset has the largest impact on this metric. The appropriateness also decreased without RL, shows the importance of data exploration.

Name	Objective evaluation				Subjective evaluation	
	Global CCA	CCA for each sequence	FGD ↓	Diversity ↑	Human-likeness	Appropriateness
Ground Truth	1.000	1.00 ± 0.00	0.0	10.03	4.22 ± 0.11	4.22 ± 0.11
Ours	**0.988**	**0.95 ± 0.02**	3.850	**7.039**	3.80 ± 0.11	**3.42 ± 0.14**
- RL	0.987	0.94 ± 0.03	**3.132**	7.008	**3.82 ± 0.11**	3.24 ± 0.16
- RL - VQVAE	0.987	0.94 ± 0.03	3.568	6.971	3.79 ± 0.11	3.33 ± 0.12
- Skeleton A	0.972	0.94 ± 0.03	13.76	4.882	3.54 ± 0.12	3.00 ± 0.13
- Skeleton B	0.965	**0.95 ± 0.03**	12.45	5.566	3.59 ± 0.13	3.09 ± 0.13

3.4 Diverse, Controllable, and Stylized Gesture Generation

The **intensity of the stylization** can be controlled by the value of γ. Due to the diffusion model architecture, different noisy gesture and different seed gesture could **generate different gestures** even for the same speech and style. We can have a **high level of control** over speech-driven gestures at any time with the specified upper body code.



Reference

[1] Skeleton-Aware Networks for Deep Motion Retargeting
[2] DiffuseStyleGesture: Stylized Audio-Driven Co-Speech Gesture Generation with Diffusion Models
[3] Bailando: 3D dance generation via Actor-Critic GPT with Choreographic Memory
[4] Edge: editable dance generation from music

Project page