UnifiedGesture: A Unified Gesture Synthesis Model for Multiple Skeletons

Sicheng Yang,1, Zilin Wang*,1, Zhiyong Wu1,4, Minglei Li2, Zhensong Zhang3, Qiaochu Huang 1, Lei Hao3, Songcen Xu3, Xiaofei Wu3, Changpeng Yang2, Zonghong Dai2*

[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

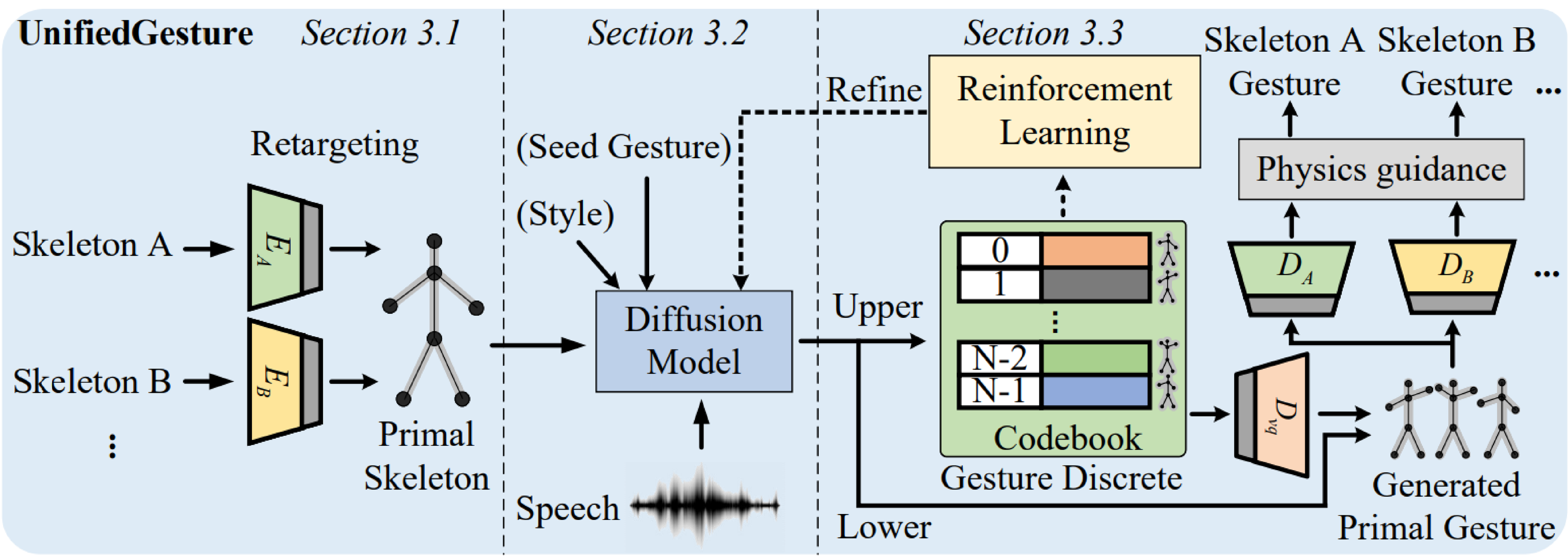
1.1 Motivation

- 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 skeletons
- ✓ Present a temporally aware attention-based diffusion model on the primal skeleton for co-speech gesture generation
- ✓ Incorporation reinforcement learning, VQVAE and IK 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



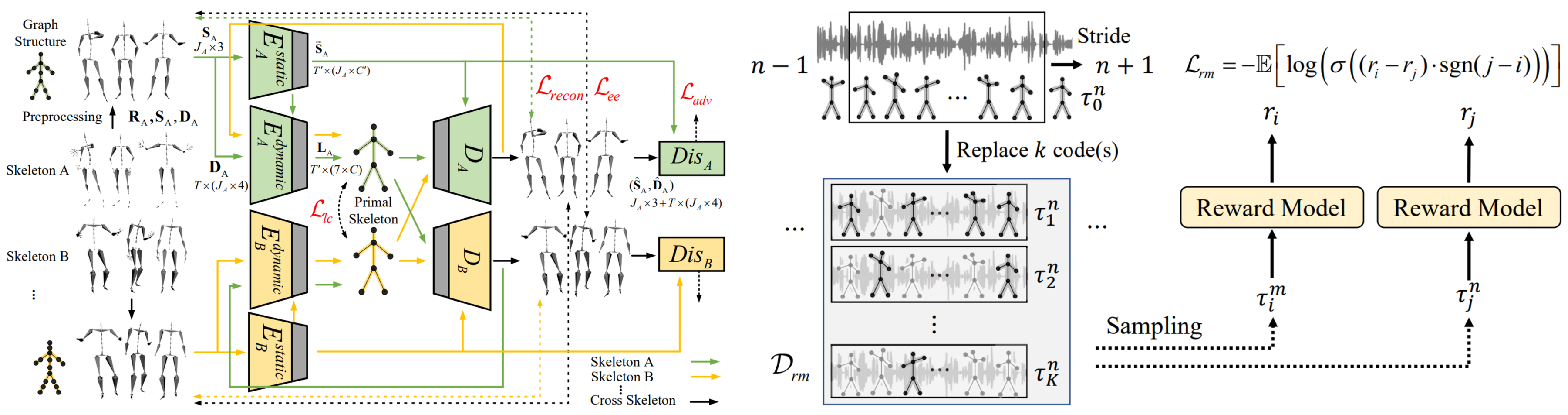
2.1 Multiple Skeletons Retargeting Network

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

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

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

According to a variance schedule $\beta_1, \beta_2, \ldots, \beta_{T_d}$ $(0 < \beta_1 < \beta_2 < \cdots < \beta_{T_d} < 1)$, T_d is the total time step),we add Gaussian noise $q(\mathbf{L}_{t_d} \mid \mathbf{L}_{t_d-1}) = \mathcal{N}(\mathbf{L}_{t_d}; \sqrt{1-\beta_{t_d}}\mathbf{L}_{t_d-1}, \beta_{t_d}\mathbf{I})$

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 $\mathbf{L}_0 = \mathrm{Denoise}(\mathbf{L}_{t_d}, t_d, c)$.

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:

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

2.3 Gesture Generation Refinement

2.3.1 Primal Gesture VQVAE

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

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

2.3.2 Reinforcement Learning Finetuning

Let the reward model R_ψ classify these trajectories with different qualities (may come from different human demonstrations with different speech) $r = R_\psi(\tau)$ to determine which trajectory is better: $\mathcal{L}_{rm} = -\mathbb{E}\left[\log\left(\sigma\left((r_i - r_j)\cdot \mathrm{sgn}(j-i)\right)\right)\right]$

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

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

3. Experiments

3.1 Experiment Preparation

- *Retargeting network.* Trinity and ZEGGS datasets. d_{re}=4, then the primal gesture is 7.5 fps. Adam optimizer, batch size of 256 for 16000 epochs.
- *Diffusion model.* Gesture data cropped to a length of N = 30 (4 seconds). AdamW optimizer, learning rate 3×10^{-5}, batch size 256, 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}, batch size of 128 for 200 epochs.

3.2 Comparison to Existing Methods

- *Human-likeness*. Our model excels beyond other top methods, matching ExampleGestures closely.
- *Gesture and speech appropriateness*. we surpass 3 baseline models, and are on par 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

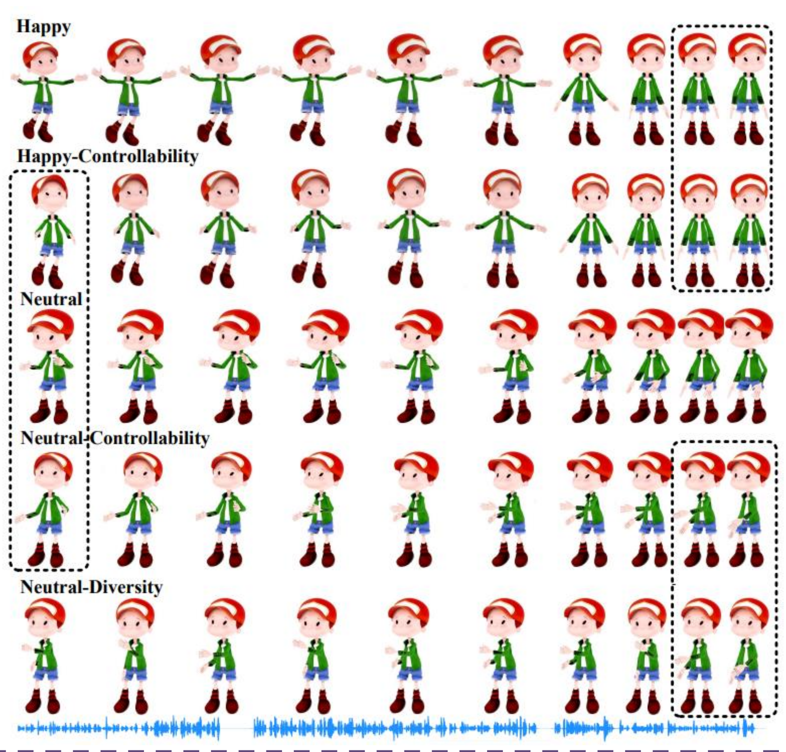
3.3 Ablation Studies

- *Human-likeness*. Significantly influenced by dataset scale, highlighting the significance of unifying gesture datasets.
- *Speech and gesture appropriateness*. Dataset size is crucial. Without RL, appropriateness decreases, highlighting 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 **stylization intensity** is regulated by γ value. Given the diffusion model, varying noise and seed gestures produce **distinct outcomes** for identical speech and style. The specified upper body code allows **precise control** over speech-driven gestures.



Reference

[1] Aberman, Kfir, et al. "Skeleton-aware networks for deep motion retargeting." ACM Transactions on Graphics (TOG) 39.4: 62-1. 2020.

[2] Yang, Sicheng, et al. "DiffuseStyleGesture: Stylized Audio-Driven Co-Speech Gesture Generation with Diffusion Models." International Joint Conference on Artificial Intelligence. 2023.

[3] Siyao, Li, et al. "Bailando: 3d dance generation by actor-critic gpt with choreographic memory." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022.

[4] Tseng, Jonathan, et al. "Edge: Editable dance generation from music." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023.

Project page