FreeTalker: Controllable Speech and Text-Driven Gesture Generation Based on Diffusion Models for Enhanced Speaker Naturalness

Sicheng Yang[1], Zunnan Xu[1], Haiwei Xue[1], Yongkang Cheng[2], Shaoli Huang[3], Mingming Gong[4,5], Zhiyong Wu[1]

[1] Tsinghua Shenzhen International Graduate School, Tsinghua University [2] Northwest A&F University

[3] Tencent AI Lab [4] University of Melbourne, [5] Mohamed bin Zayed University of Artificial Intelligence

1. Introduction

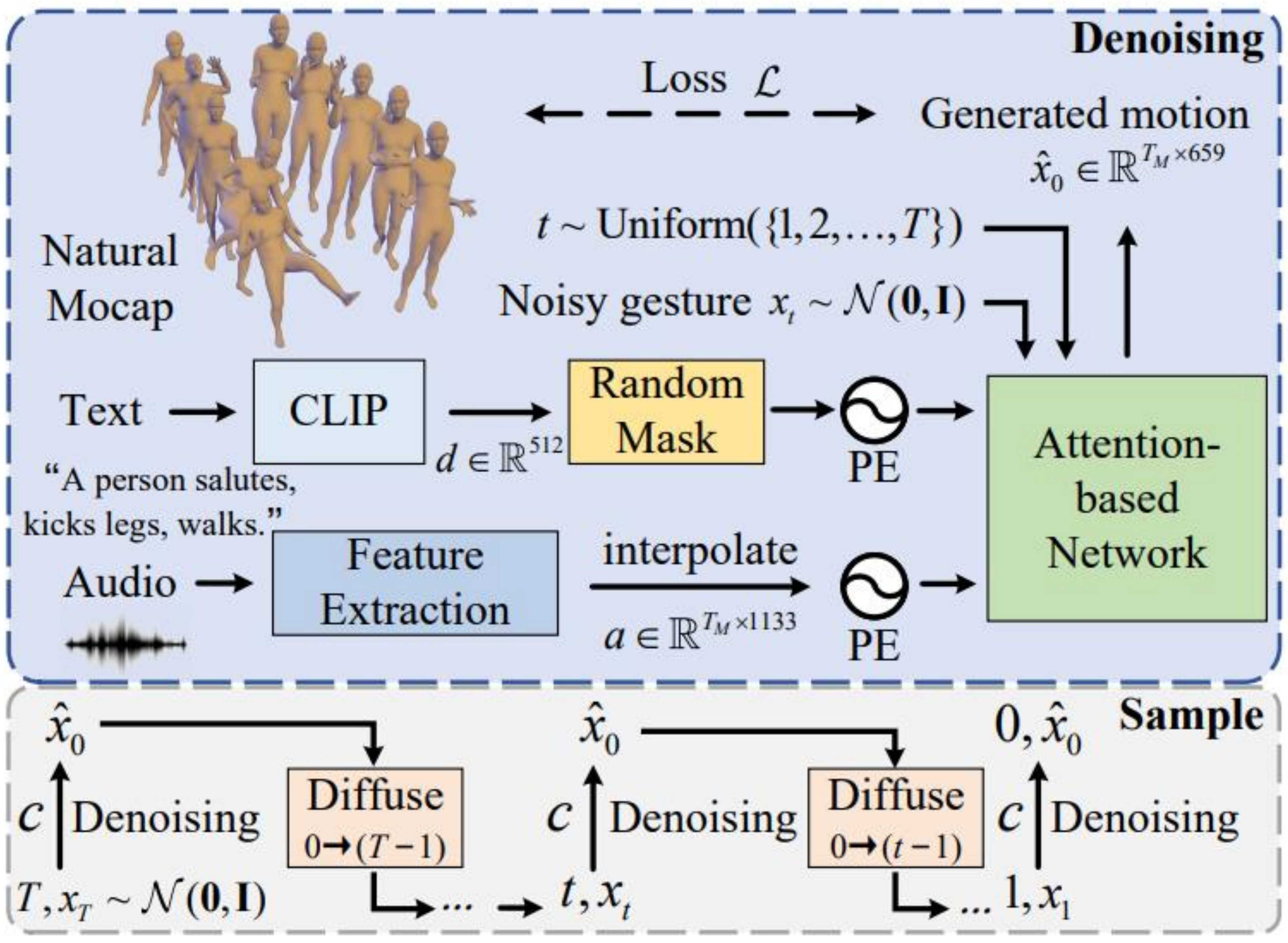
1.1 Motivation

- **Why Free Speaker Motions?**
  - IMPORTANT in virtual agents, animation, HCI
  - Including co-speech gestures and movement like *walking, pointing or interacting* — is crucial for realism and engagement
- **Limitations of Current Work**
  - Focus on co-speech gesture generation
  - Limited focus on **free** motion (spontaneous and non-spontaneous)
- **Challenges**
  - Disjointed motion representations and diverse inputs handling → multi-dataset utilization and multimodal learning
  - Long sequence and controllable motion generation

1.2 Contributions

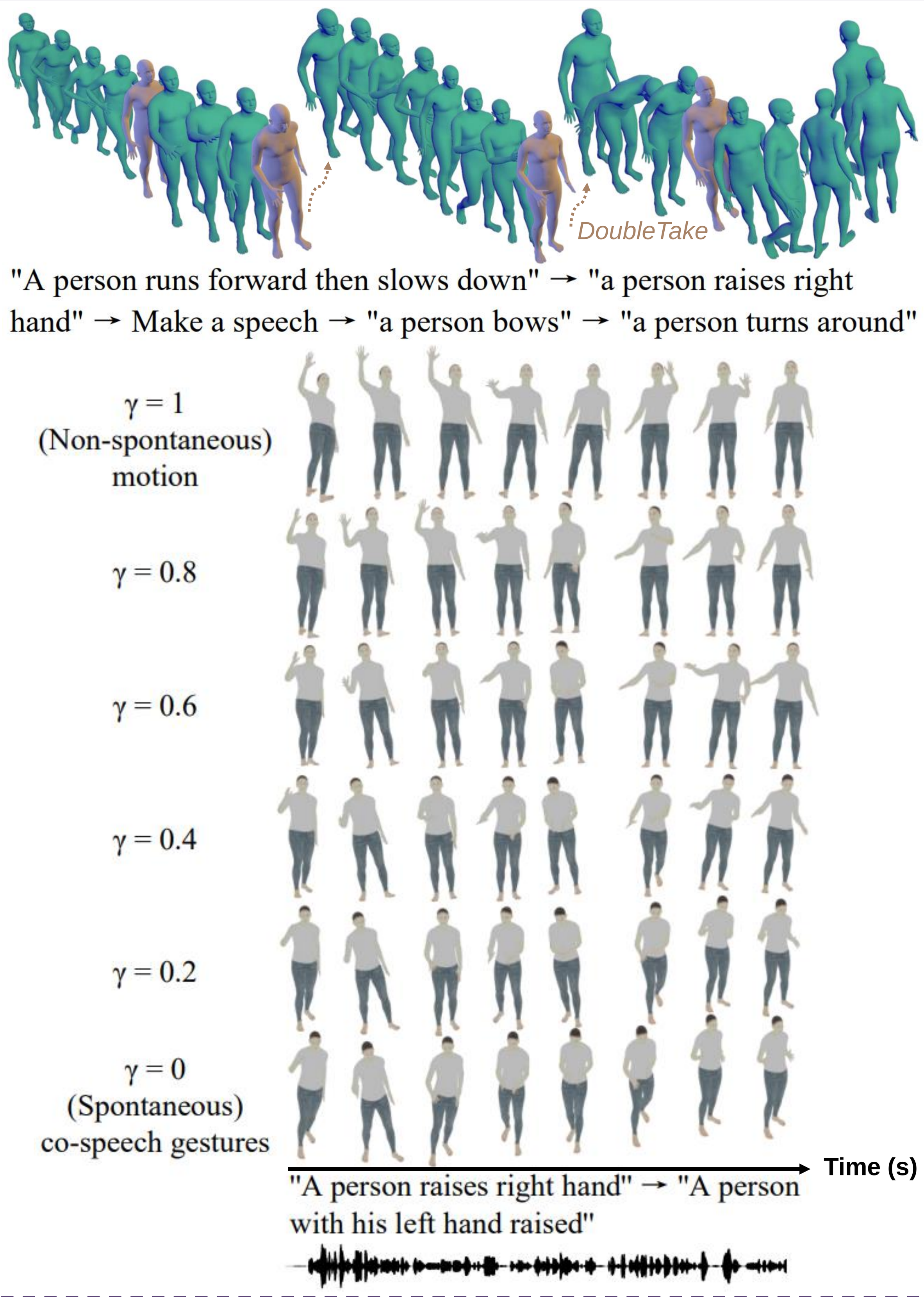
- ✓ The first framework to generate free speaker motions
- ✓ Employing classifier-free guidance and DoubleTake for controlled, flexible gesture generation
- ✓ Demonstrating increased naturalness in speaker motions

2. Methodology



- **Motion Processing**
  - **Adaptation**: converts BVH to axis-angle (SMPL-X) for detailed motion; adapts 3D positions to SMPL-X with Vposer with uniform scale and root joint translations
  - **Features**: includes root height, linear/rotational velocities, joint rotation/position/velocity, and foot contact
- **Diffusion Model for Motion Generation**
  - **Conditioning**: integrates text and audio inputs to generate motion
  - **Implementation**:
    - T steps, and initial motion is derived from a normal distribution
    - Predict clean motion $\hat{x}_0$ from noised inputs x_t, incorporating text (CLIP) and audio features (WavLM etc.) as conditions
    - Huber loss function
- **Controllable Long Motion Generation using DoubleTake**
  - **Conditioning**: uses text / audio to generate gestures, balancing inputs through a mix parameter (γ)
  - **Implementation**: blend and smooth transitions between motion segments, ensuring seamless long-duration motion generation

3. Visualization



4. Experiments

- **Datasets**: HumanML3D (text-driven) and BEAT (speech-driven)
  - **Preprocessing**: resampling to 20 FPS; HumanML3D spans 40-180 frames, texts up to 20 words; English speakers' gestures
  - **Split**: 80% train, 10% validate, 10% test; weighted sampling
  - **Normalization**: mean subtraction and standard deviation
- **Model**: T=1000, cosine schedule, 256-dimension self-attention
- **Training**: 1M steps, batch size 256, learning rate 2e-4, over 3 days on one V100 GPU

Name	Co-speech gesture generation			
	jerk →	acceleration →	FID ↓	Naturalness ↑
Natural Mocap	135.36 ± 58.61	12.39 ± 11.79	-	-
DiffuseStyleGesture	206.52 ± 83.65	5.68 ± 2.19	0.008	49%
MDM	-	-	-	-
Ours*	245.78 ± 108.27	6.03 ± 2.55	0.139	40%*

Name	Motion Generation			Free-motion	
	SSIM ↑	FID ↓	Naturalness ↑	FID ↓	Naturalness ↑
Natural Mocap	-	-	-	-	-
DiffuseStyleGesture	-	-	-	-	-
MDM	0.386	0.050	53%	-	-
Ours*	0.457	0.226	24%*	0.139	-

- **Results**
  - **Objective**: competitive results of our method with baselines
  - **Subjective**: user study on *naturalness*, 25 participants; competitive performance in comparison to baselines, suggesting improvements with an expanded motion database

References

[1] Tevet G, Raab S, Gordon B, et al. Human motion diffusion model//The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023.

[2] Yang S, Wu Z, Li M, et al. Diffusestylegesture: Stylized audio-driven co-speech gesture generation with diffusion models[C]//Proceedings of the 32nd International Joint Conference on Artificial Intelligence, IJCAI 2023, Macao, S.A.R, 19th-25th August 2023.

Project page