

InteracTalker: Prompt-Based Human-Object Interaction with Co-Speech Gesture Generation

Anonymous WACV Algorithms Track submission

Paper ID 2882

A. 3D Human Motion Representation

Our 3D human motion is represented using SMPL pose parameters, a standard practice in recent work. The orientation of each of the 22 joints (including the root joint) is encoded using 6D rotations [7] to avoid gimbal lock and ensure continuity. Instead of using absolute global translations, we represent the global body movement as the difference between consecutive frames [1, 5]. This approach helps to generate smoother motions by focusing on local displacement rather than absolute position, which can suffer from drift. Consequently, each motion frame is represented as a 135-dimensional vector, comprising the 6D rotations for all 22 joints and a 3D vector for the consecutive global translation. A full motion is a sequence of these 135-dimensional vectors. Before training, these features are normalized using the mean and variance of the training set, a standard procedure to stabilize the learning process.

B. Adaptation Module Encoders

This section provides a detailed overview of the two primary encoders used to generate conditioning signals for our adaptation modules. The architecture of these encoders is illustrated in Figure 1.

As shown in Figure 1(a), the Speech Content Encoder is designed to process the multimodal inputs required for co-speech gesture generation. It takes both speech audio and its corresponding transcript, processes them to extract distinct features, and then concatenates these to produce a joint audio-text feature vector. This feature vector subsequently serves as the conditioning signal for the Co-speech Gesture Adaptation Branch.

Figure 1(b) details the BPS Feature Generator, which is responsible for deriving object-specific conditioning signals. This encoder takes as input the current noisy body pose and the geometry of the target object. It extracts two types of features: object geometry features from Basis Point Sets (BPS) and human-object interaction features. These are then merged to create a comprehensive feature vector that guides the Interaction-Aware Adaptation Branch, en-

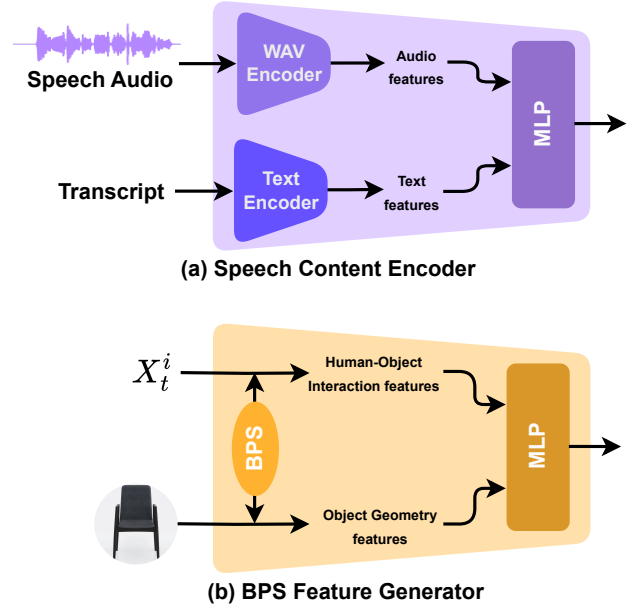


Figure 1. Architecture of InteracTalker’s Adaptation Encoders. (a) The Speech Content Encoder extracts joint audio-text features from speech audio and a transcript for co-speech gesture generation. (b) The BPS Feature Generator derives object geometry and human-object interaction features from the body pose and a target object to guide object-aware motion.

abling our model to generate motions that are highly responsive to a target object’s properties.

C. Experimental Setup

Our experimental setup follows a three-stage training process, leveraging a modular approach to build the InteracTalker framework.

First, the base Motion Diffusion Model (MDM) was pre-trained using the HumanML3D [3] and SAMP [4] datasets, which do not contain object interactions. This training was performed for 500k steps with a batch size of 200. Following this, the MDM was frozen to serve as a stable motion

prior for the subsequent training of our adaptation branches.

Next, the interaction-aware branch was trained for 110k steps with a batch size of 64. This stage utilized the SAMP [4] dataset, augmented with motion-aligned objects from the 3D-FRONT [2] dataset, along with their corresponding text prompts.

Finally, the co-speech aware branch was trained separately for 600k steps with a batch size of 64 using the BEATX [6] dataset. This allowed it to specialize in the fine-grained dynamics of co-speech gestures.

For all training stages, we used the AdamW optimizer with a learning rate of 10^{-4} . All experiments were conducted on a single NVIDIA GeForce RTX 2080 Ti GPU, with the total training time spanning approximately 5 days. The number of diffusion steps was set to 1000 for both training and inference.

During inference, the conditioning signals from both adaptation branches are synergistically injected into the pre-trained and frozen MDM, enabling the generation of comprehensive motions that account for both co-speech dynamics and object interactions simultaneously.

D. Qualitative Results of InteracTalker

This section provides a visual demonstration of our method's performance in generating realistic and physically plausible human-object interactions. Close-up views of the interaction regions in the generated motions are presented in Figure 2, visually demonstrating that our approach effectively maintains the necessary physical contact for realistic interaction while rigorously minimizing penetration with the target object.

In addition to Figure 2 presented below, a supplementary video is provided to offer a dynamic view of our qualitative results. A 'README' file is also included with the video, which details the specific content and examples shown.



Figure 2. Qualitative Results: Physically Plausible Human-Object Interaction by InteracTalker. This figure demonstrates our method's superior ability to generate realistic human-object interactions. Observe how our approach consistently maintains necessary physical contact with the target object (e.g., foot on ground, hand on chair) while achieving minimal penetration, a key challenge in interaction synthesis. These results visually corroborate our low quantitative penetration metrics.

E. User Study

As the first work to address the complex task of generating co-speech gestures that simultaneously account for detailed object interactions, we conducted a user study to assess the perceived plausibility and realism of our method. We compared our full InteracTalker framework with a naive concatenation (Concat) approach, which combines independently generated upper-body co-speech gestures and lower-body object-interaction motions. We recruited around 64 participants with minimal or no prior expertise in the domain to ensure an unbiased evaluation based on visual clarity and naturalness. Participants were presented with pairs of videos from both our method and the naive approach. For each pair, they answered questions about different aspects of motion quality. The full evaluation interface is provided in Figure 3. The user study results, demonstrate a strong preference for our InteracTalker framework (preferred 75%). Our method achieved a significantly higher average rating for motion realism and was preferred by a large majority of participants across all other metrics. This confirms that a unified, integrated solution is critical for generating realistic motions, as the naive approach consistently produced motions with undesirable artifacts such as self- and object-penetrations.

References

- [1] Nikos Athanasiou, Alpár Cseke, Markos Diomataris, Michael J Black, and Gül Varol. Motionfix: Text-driven 3d human motion editing. In *SIGGRAPH Asia 2024 Conference Papers*, pages 1–11, 2024. 1
- [2] Huan Fu, Bowen Cai, Lin Gao, Ling-Xiao Zhang, Jiaming Wang, Cao Li, Qixun Zeng, Chengyue Sun, Rongfei Jia, Binqiang Zhao, et al. 3d-front: 3d furnished rooms with layouts and semantics. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10933–10942, 2021. 2
- [3] Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating diverse and natural 3d human motions from text. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5152–5161, 2022. 1
- [4] Mohamed Hassan, Duygu Ceylan, Ruben Villegas, Jun Saito, Jimei Yang, Yi Zhou, and Michael J Black. Stochastic scene-aware motion prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11374–11384, 2021. 1, 2
- [5] Daniel Holden, Jun Saito, and Taku Komura. A deep learning framework for character motion synthesis and editing. *ACM Transactions on Graphics (ToG)*, 35(4):1–11, 2016. 1
- [6] Haiyang Liu, Zihao Zhu, Giorgio Becherini, Yichen Peng, Mingyang Su, You Zhou, Xuefei Zhe, Naoya Iwamoto, Bo Zheng, and Michael J Black. Emage: Towards unified holistic co-speech gesture generation via expressive masked audio gesture modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1144–1154, 2024. 2

User Study

We are conducting a user study to evaluate the quality of computer-generated human motion in scenarios involving **human-object interaction** (e.g., sitting on a chair) combined with **co-speech gestures** (e.g., hand and body movements aligned with speech). The goal is to understand which motion generation method produces more natural, realistic, and coherent results.

What You Will See

You will be shown **pair of short video clips (Video A and Video B)**. Each pair depicts the same scenario but generated using two different methods. The order of videos is randomized.

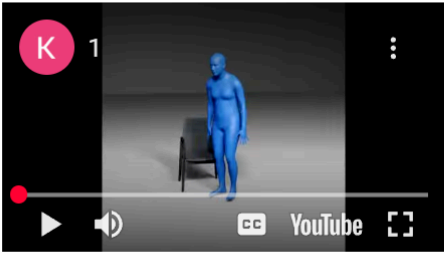
Your Task

- For each video pair, you will be asked to:
- Compare **naturalness and realism** of body motion.
 - Evaluate the quality of **human-object interaction**.
 - Judge how well **gestures align with speech**.
 - Identify if there are noticeable **penetrations** (self-penetration or body-object penetration).
 - Indicate your **overall preference** between the two videos.

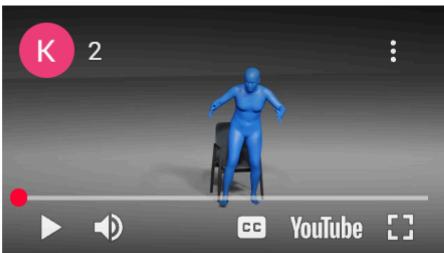
This study includes **speech audio** along with the animated videos.

- Please make sure your **speakers are on** or you are wearing **headphones**.
- The audio is important to evaluate how well the generated gestures and body motions align with the speech.

Video A



Video B



Rate the **realism** of the motion in **video A** on a scale of 1-5. *

	1	2	3	4	5	
Very Unrealistic	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Very Realistic

Rate the **realism** of the motion in **video B** on a scale of 1-5. *

	1	2	3	4	5	
Very Unrealistic	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Very Realistic

Which video looks more **natural and realistic**? *

☐ Video A

☐ Video B

Which video shows **better interaction with object**? *

☐ Video A

☐ Video B

Which video better aligns **gestures and body movements with speech**? *

☐ Video A

☐ Video B

In which video did the human body show **fewer unrealistic penetrations with itself** (e.g., arms going inside the body). *

☐ Video A

☐ Video B

In which video did the human body show **fewer unrealistic penetrations with the object** (e.g., sitting inside the chair instead of on it, hand clipping through)? *

☐ Video A

☐ Video B

Overall, which video would you **prefer to watch in a real application**? *

☐ Video A

☐ Video B

Figure 3. Interface used in conducting user study.

- 138 [7] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao
139 Li. On the continuity of rotation representations in neural net-
140 works. In *Proceedings of the IEEE/CVF conference on com-*
141 *puter vision and pattern recognition*, pages 5745–5753, 2019.
142 1