

bunGesture : Diffusion Gesture Generation Model for Vietnamese language

Thanh Hoang-Minh

hmthanhgm@gmail.com

Department of Information Technology, VNUHCM -
University of Science
Ho Chi Minh City, District 5, Vietnam

Ngoc Ly-Quoc

lqngoc@fit.hcmus.edu.vn

Department of Information Technology, VNUHCM -
University of Science
Ho Chi Minh City, District 5, Vietnam

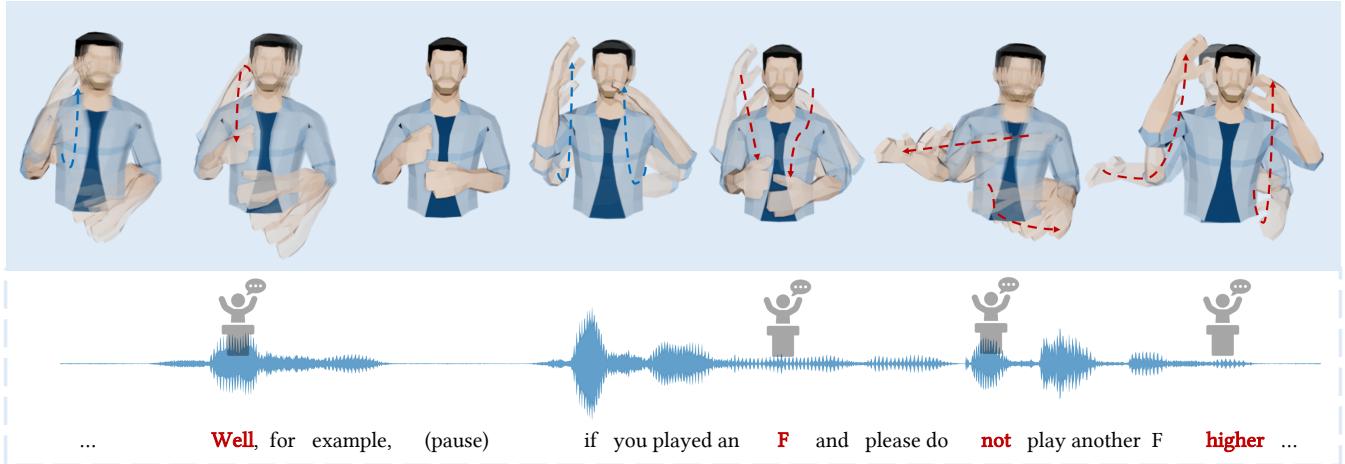


Figure 1: Gesture results automatically synthesized by our system for a beat-rich TED talk clip. The red words represent beats, and the red arrows indicate the movements of corresponding beat gestures.

ABSTRACT

Gesture generation is a pivotal aspect of multimodal communication systems, enhancing the naturalness and contextuality of human-machine interactions. While significant progress has been made in this field for various languages, gesture generation tailored specifically for Vietnamese remains underexplored. This paper introduces , a novel diffusion-based gesture generation model designed for the Vietnamese language. Unlike traditional approaches relying on English-centric models, bunGesture leverages Vietnamese-specific tools and techniques.

Key to bunGesture is its adoption of diffusion-based gesture generation, integrating advanced Vietnamese language models like

PhoBERT and the VnCoreNLP toolkit. Additionally, bunGesture employs a fine-tuned Vietnamese version of WaveLM, trained on a custom Vietnamese audio dataset.

This work emphasizes the importance of language-specific adaptations in gesture generation systems and opens avenues for similar research in other languages.

CCS CONCEPTS

- Computing methodologies → Animation; Natural language processing; Neural networks.

KEYWORDS

non-verbal behavior, multi-modality, virtual agents, co-speech gesture synthesis, character animation, neural generative model

ACM Reference Format:

Thanh Hoang-Minh and Ngoc Ly-Quoc. 2024. bunGesture : Diffusion Gesture Generation Model for Vietnamese language. In *Proceedings of ICMI 2024*. ACM, New York, NY, USA, 19 pages. <https://doi.org/XXXXXX.XXXXXXX>

1 INTRODUCTION

Gesturing is an important part of speaking. It adds emphasis and clarity to a speech and conveys essential non-verbal information that makes the speech lively and persuasive [Burgoon et al. 1990]. There are rich demands for high-quality 3D gesture animation in many industries, such as games, films, and digital humans. However,

Authors' addresses: Thanh Hoang-Minh, hmthanhgm@gmail.com, Department of Information Technology, VNUHCM - University of Science, 227 Nguyen Van Cu Street, Ward 4, Ho Chi Minh City, District 5, Vietnam, 70000; Ngoc Ly-Quoc, lqngoc@fit.hcmus.edu.vn, Department of Information Technology, VNUHCM - University of Science, 227 Nguyen Van Cu Street, Ward 4, Ho Chi Minh City, District 5, Vietnam, 70000.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

ICMI 2024, November 04–08, 2024, San José, Costa Rica

© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM. ACM ISBN 978-1-4503-XXXX-X/18/06
<https://doi.org/XXXXXX.XXXXXXX>

the difficulties in reproducing the complex yet subtle harmony between vocalization and body movement make synthesizing natural-looking co-speech gestures remain a long-standing and challenging task.

Gestures are grouped into six categories by linguists [Ekman and Friesen 1969; McNeill 1992]—adaptors, emblems, deictics, iconics, metaphors, and beats. Among them, the beat gestures are rhythmic movements that bear no apparent relation to speech semantics [Kipp 2004] but serve meta-narrative functions [McNeill 1992] that are crucial to rhythmic harmony between speech and gestures. Generating realistic beat gestures requires modelling the relation between the gestural beats and the verbal stresses. However, it has been observed that these two modalities are not synchronized in a strict rhythmic sense [McClave 1994], making it difficult to learn their temporal connection directly from data using an end-to-end method [Bhattacharya et al. 2021a; Kucherenko et al. 2020; Yoon et al. 2020].

Gestures are associated with different levels of speech information [McNeill 1992]. For example, an emblem gesture such as *thumbs-up* usually accompanies high-level semantics like *good* or *great*, while a beat gesture commonly comes with low-level acoustic emphasis. Many previous studies use only the features extracted at the last layer of an audio encoder to synthesize gestures [Alexanderson et al. 2020; Bhattacharya et al. 2021a; Kucherenko et al. 2020; Qian et al. 2021; Yoon et al. 2020]. This setup, however, may in effect encourage the encoder to mix the speech information from multiple levels into the same feature, causing ambiguity and increasing the difficulty in mining clear rhythmic and semantic cues.

In this paper, we focus on generating co-speech upper-body gestures that can accompany a broad range of speech content—from a single sentence to a public speech, aiming at achieving convincing results both on the rhythm and semantics. Our first observation is that gesturing can be considered as a special form of dancing under changing beats. We develop a rhythm-based canonicalization and generation framework to deal with the challenge of generating synchronized gestures to the speech, which segments the speech into short clips at audio beats, normalizes these clips into canonical blocks of the same length, generates gestures for every block, and aligns the generated motion to the rhythm of the speech. This framework, which is partially inspired by recent research in dance generation [Aristidou et al. 2022], provides the gesture model with an explicit hint of the rhythm, allowing the model to learn the pattern of gestural beats within a rhythmic block efficiently. Both the quantitative evaluation with a novel rhythmic metric and the qualitative evaluation with user studies show that the gestures generated by this pipeline exhibit natural synchronization to the speech.

As indicated in linguistics literature [Kipp 2004; Neff et al. 2008; Webb 1996], gestures used in everyday conversation can be broken down into a limited number of semantic units with different motion variations. We assume that these semantic units, usually referred to as *lexemes*, relate to the high-level features of speech audio, while the motion variations are determined by the low-level audio features. We thus disentangle high- and low-level features from different layers of an audio encoder and learn the mappings between them and the gesture lexemes and the motion variations,

respectively. Experiments demonstrate that this mechanism successfully disentangles multi-level features of both the speech and motion and synthesizes semantics-matching and stylized gestures.

In summary, our main contributions in this paper are:

- We present a novel rhythm- and semantics-aware co-speech gesture synthesis system that generates natural-looking gestures. To the best of our knowledge, this is the first neural system that explicitly models both the rhythmic and semantic relations between speech and gestures.
- We develop a robust rhythm-based segmentation pipeline to ensure the temporal coherence between speech and gestures, which we find is crucial to achieving rhythmic gestures.
- We devise an effective mechanism to relate the disentangled multi-level features of both speech and motion, which enables generating gestures with convincing semantics.

2 RELATED WORK

2.1 Data-driven Human Motion Synthesis

Traditional human motion synthesis frameworks often rely on concatenative approaches such as *motion graph* [Kovar et al. 2002]. Recently, learning-based methods with neural networks have been widely applied to this area to generate high-quality and interactive motions, using models ranging from feed-forward network [Holden et al. 2017; Starke et al. 2022] to dedicated generative models [Henter et al. 2020; Ling et al. 2020]. Dealing with the one-to-many issue where a variety of motions can correspond to the same input or control signal is often a challenge for these learning-based approaches. Previous systems often employ additional conditions, such as contacts [Starke et al. 2020] or phase indices [Holden et al. 2017; Starke et al. 2022], to deal with this problem. Closer to the gesture domain is the speech-driven head motion synthesis, where conditional GANs [Sadoughi and Busso 2018], and conditional VAEs [Greenwood et al. 2017] have been used.

2.1.1 Music-driven Dance Synthesis. Among the general motion synthesis tasks, music-driven dance generation addresses a similar problem to the co-speech gesture synthesis, where the complex temporal relation between two different modalities needs to be modeled accurately. Both motion graph-based methods [Chen et al. 2021; Kim et al. 2006] and learning-based approaches [Li et al. 2021b; Siyao et al. 2022; Valle-Pérez et al. 2021] have been adopted and successfully achieved impressive generation results. To deal with the synchronization between the dance and music, Chen et al. [2021] develop a manually labeled rhythm signature to represent beat patterns and ensures the rhythm signatures of the generated dance match the music. Aristidou et al. [2022] segment the dance into blocks at music onsets, convert each block into a motion motif [Aristidou et al. 2018] that defines a specific cluster of motions, and use the motion motif to guide the synthesis of dance at the block level. Siyao et al. [2022] employ a reinforcement learning scheme to improve the rhythmic performance of the generator using a reward function encouraging beat alignment. Our rhythm-based segmentation and canonicalization framework is partially inspired by [Aristidou et al. 2022]. Similar to [Aristidou et al. 2022], we also segment the gestures into clips at audio beats but learn a high-level representation for each clip via the vector quantization scheme

[Oord et al. 2017] instead of the K-means clustering. Moreover, our framework generates gestures in blocks of motion and denormalizes the generated motion blocks to match the rhythm of the speech. In contrast, Aristidou et al. [2022] synthesize dance sequences in frames conditioned on the corresponding motion motifs.

2.2 Co-speech Gesture Synthesis

The most primitive approach used to generate human non-verbal behaviors is to animate an artificial agent using the retargeted motion capture data. This kind of approach is widely used in commercial systems (e.g., films and games) because of its high-quality motion performance. However, it is not suitable for creating interactive content that cannot be prepared beforehand. Generating co-speech gestures according to an arbitrary input has been a long-standing research topic. Previous studies can be roughly categorized into two groups, i.e., rule-based and data-driven methods.

2.2.1 Rule-based Method. The idea of the rule-based approach is to collect a set of gesture units and design specific rules that map a speech to a sequence of gesture units [Cassell et al. 2004; Huang and Mutlu 2012; Kipp 2004; Softbank 2018]. Wagner et al. [2014] have an excellent review of these methods. The results of the rule-based methods are generally highly explainable and controllable. However, the gesture units and rules typically have to be created manually, which can be costly and inefficient for complex systems.

2.2.2 Data-driven Method. Early research in data-driven method learns the rules embedded in data and combines them with pre-defined animation units to generate new gestures. For example, Kopp et al. [2006]; Levine et al. [2010] use probabilistic models to build correspondence between speech and gestures. Neff et al. [2008] build a statistical model to learn the personal style of each speaker. The model is combined with the input text tagged with the theme, utterance focus, and rheme to generate gesture scripts, which are then mapped to a sequence of gestures selected from an animation lexicon. Chiu et al. [2015] train a neural classification model to select a proper gesture unit based on the speech input. More recent research has started to take advantage of deep learning and trains end-to-end models using raw gesture data directly, which frees the manual efforts of designing the gesture lexicon and mapping rules. Gestures can be synthesized using deterministic models such as multilayer perceptron (MLP) [Kucherenko et al. 2020], recurrent neural networks [Bhattacharya et al. 2021a; Hasegawa et al. 2018; Liu et al. 2022; Yoon et al. 2020, 2019], convolutional networks [Habibie et al. 2021], and transformers [Bhattacharya et al. 2021b], or by learning generative models such as normalizing flow [Alexanderson et al. 2020], VAEs [Li et al. 2021a; Xu et al. 2022], and learnable noise codes [Qian et al. 2021]. Our method is also a data-driven framework. We learn the motion generator and the mapping between the speech and gestures from data using a combined network structure of the vector quantized variational autoencoder (VQ-VAE) [Oord et al. 2017] and LSTM. To capture the rhythmic and semantic correspondences between the speech and gestures, we propose a multi-stage architecture that explicitly models the rhythm and semantics in different stages. An earlier system proposed by Kucherenko et al. [2021b] shares a similar high-level architectural design to our framework. However, there are two key

differences: (a) our method is essentially an unsupervised learning approach, which learns the gesture lexeme, style code, and the generator directly from the data without detailed annotations; and (b) our system employs an explicit beat-based segmentation scheme which is shown to be effective in ensuring temporal coherence between the speech and the gesture.

2.3 Multi-Modal Data Processing

Co-speech gesture generation is a cross-modal process involving audio, text, motion, and other information related to the speaker and the content of the speech. The representation and alignment of each modality are essential for high-quality results [Baltrušaitis et al. 2019]. Mel-spectrogram and MFCC acoustic features are commonly used as audio features [Alexanderson et al. 2020; Kucherenko et al. 2020; Qian et al. 2021], typically resampled into the same framerate of the motion. For the text features, pre-trained language models like BERT [Devlin et al. 2019; Kucherenko et al. 2020] and FastText [Bojanowski et al. 2017; Yoon et al. 2020] have been used to encode text transcripts into frame-wise latent codes, where paddings, fillers, or empty words are inserted into a sentence to make the word sequence the same length as the motion [Kucherenko et al. 2020; Yoon et al. 2020]. Speaker’s style and emotions can also be encoded by learnable latent codes [Bhattacharya et al. 2021a; Yoon et al. 2020] and are resampled or padded to match the length of the speech. In this work, we employ a pre-trained speech model to extract audio features and fine-tune it using a contrastive learning strategy. We also utilize a BERT-based model to vectorize the text. These multi-modal data are then aligned explicitly using the standard approaches discussed above. Notably, a concurrent study [Liu et al. 2022] also extracts audio features using contrastive learning. Their framework considers the learning of the audio features as a part of the training of the gesture generator. Instead, our framework trains the audio encoder in a separate pre-training stage using only the audio data.

2.4 Evaluation of Motion Synthesis Models

Evaluating the generated co-speech gestures is often difficult because the motion quality is a very subjective concept. Previous works have proposed several evaluation criteria. Wolfert et al. [2022] have made a comprehensive review of them. User studies are widely adopted to evaluate different aspects of motion quality, such as human-likeness and speech-gesture matching [Alexanderson et al. 2020; Kucherenko et al. 2020; Yoon et al. 2020], but can be expensive and hard to exclude uncontrolled factors. The absolute difference of joint positions or other motion features, such as velocity and acceleration between a reconstructed motion and the ground truth, is used by several works as an objective metric [Ginosar et al. 2019; Joo et al. 2019; Kucherenko et al. 2019]. However, this metric is not suitable for evaluating motions that are natural but not the same as the reference. Fréchet Inception Distance (FID) [Heusel et al. 2017] is a widely used criterion in image generation tasks that measures the difference between the distributions of the dataset and generated samples in the latent space. It successfully reflects the perceptual quality of generated samples. Similarly, Yoon et al. [2020] and Qian et al. [2021] propose Fréchet Gesture Distance (FGD) and Fréchet Template Distance (FTD) metrics, respectively.

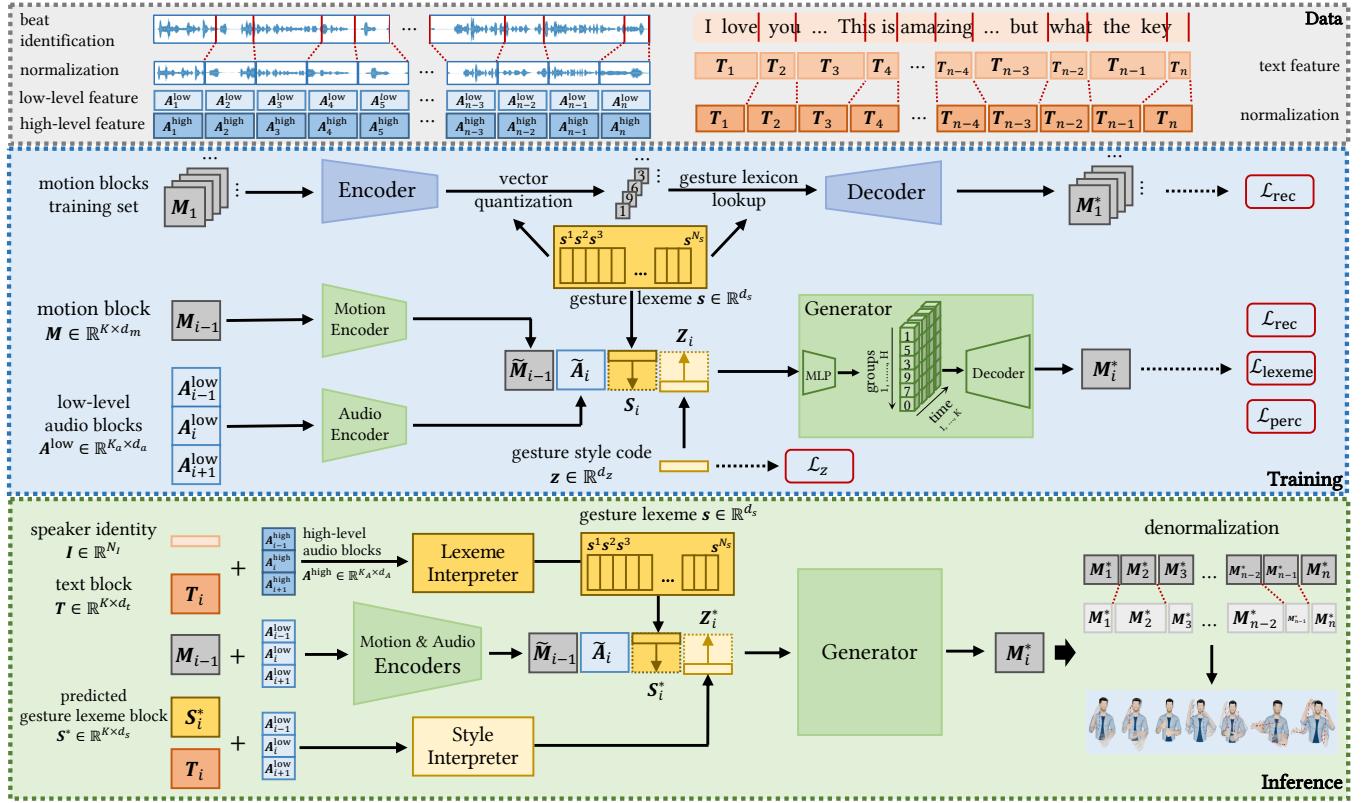


Figure 2: Our system is composed of three core components: (a) the *data* module preprocesses a speech, segments it into normalized blocks based on the beats, and extracts speech features from these blocks; (b) the *training* module learns a gesture lexicon from the normalized motion blocks and trains the generator to synthesize gesture sequences, conditioned on the gesture lexemes, the style codes, as well as the features of previous motion blocks and adjacent speech blocks; and (c) the *inference* module employs interpreters to transfer the speech features to gesture lexemes and style codes, which are then used by the learned generator to predict future gestures.

These metrics measure the perceptual quality of generated gestures. In this paper, we compare our framework with several baseline methods using both user studies and objective metrics like FGD. We further propose a simple but effective rhythmic metric to measure the percentage of matched beats by dynamically adjusting the matching threshold, which provides a more informative picture of the rhythm performance.

3 SYSTEM OVERVIEW

Our goal is to synthesize realistic co-speech upper-body gestures that match a given speech context both temporally and semantically. To achieve this goal, we build a system using neural networks that takes speech audio as input and generates gesture sequences accordingly. Additional speech modalities, such as text and speaker identity, will also be considered by the system when available to enhance semantic coherence and generate stylized gestures.

A gesture motion consists of a sequence of gesture units, which can be further broken down into a number of gesture phases that align with intonational units, such as pitch accents or stressed syllables [Kendon 2004; Loehr 2012]. The action in each of these gesture

phases is typically a specific movement such as lifting a hand, holding an arm at a position, or moving both arms down together, which is often referred to as a *gesture lexeme* by linguists [Kipp 2004; Neff et al. 2008; Webb 1996]. It is also revealed in the literature that there are only a limited number of lexemes used in everyday conversation. These lexemes form a *gesture lexicon*. A typical speaker may only use a subset of this lexicon and apply slight variations to the motion. We assume such variations cannot be inferred directly from the speech but can be characterized by some latent variables, which we refer to as the *gesture style codes*. Our system then generates gestures in a hierarchical order. It first determines the sequence of gesture lexemes and style codes and then generates gestural moves based on these motion-related features and other speech modalities.

Our system processes the input speech in a block-wise manner. Considering the temporal and structural synchrony between the gesture and the speech, we leverage a segmentation that aligns with the rhythm of the speech to ensure temporal coherence between the two modalities. Specifically, our system extracts beats from the input speech based on audio onsets and segments the speech into short clips at every beat. These clips are then time-scaled and converted into normalized blocks with the same length. We extract



Figure 3: The character model used in our system.

features at multiple levels for each block, where the high-level features are translated into a gesture lexeme, and the low-level features determine the style code. The generated gesture motions are then denormalized to match the length of the input speech.

As illustrated in Figure 2, our system consists of three core components: (a) the *data* module preprocesses a speech, segments it into normalized blocks based on the beats, and extracts speech features from these blocks; (b) the *training* module learns a gesture lexicon from the normalized motion blocks and trains the generator to synthesize gesture sequences, conditioned on the gesture lexemes, the style codes, as well as the features of previous motion blocks and adjacent speech blocks; and (c) the *inference* module employs interpreters to transfer the speech features to gesture lexemes and style codes, which are then used by the learned generator to predict future gestures.

We train our system on a speech-gesture dataset with accompanying text and speaker identity (ID). The gesture lexicon is constructed using unsupervised learning based on a vector quantized variational autoencoder (VQ-VAE) [Oord et al. 2017]. The generator is trained as an autoregressive encoder-decoder network, where we use an LSTM-based decoder combined with a vector quantized encoder to generate gesture motions. We train two separate interpreters to translate speech features into gesture lexemes and style codes, respectively. These interpreters can work with only the audio features and can be retrained to accept other speech modalities. In the following sections, we will provide details about these components and how they are trained in our system.

4 DATA PREPARATION

The *data* module of our system preprocesses an input speech, segments it into uniform blocks based on speech rhythms, and extracts features that will be used to generate co-speech gestures. In this section, we first introduce the representations of different speech modalities and gesture motion and then describe details of the data preprocessing.

4.1 Representation of Speech Modalities

4.1.1 Motion Representation. We focus on upper-body gestures in this work. Our system employs a character model consisting of 16 upper-body joints, including a rotational root, as shown in Figure 3. A gesture pose is then represented as a list of joint rotations, parameterized using the exponential map, in the hierarchical order. We use $\mathbf{m}_k \in \mathbb{R}^{d_m}$ to represent the gesture pose at frame k , and a

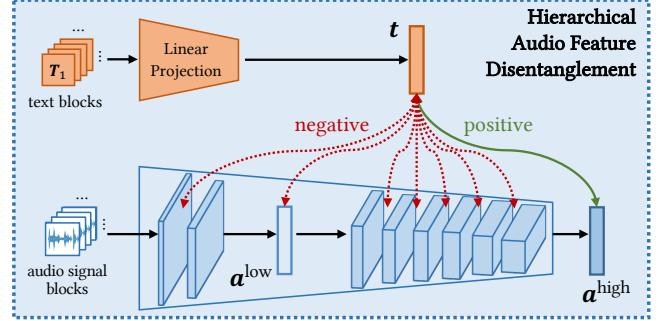


Figure 4: A contrastive learning task is performed to disentangle multi-level audio features. We use the text feature t as the anchor of this learning. The highest-level audio feature a^{high} is considered as the positive sample, while the features of the lower levels are all treated as negative samples. After training, both a^{high} and the feature extracted at the second level, a^{low} , are used for gesture generation.

clip of gestures is represented collectively as $\mathbf{M} = \{\mathbf{m}_1, \dots, \mathbf{m}_K\}$, where K is the number of frames. We retarget training motions onto this model by copying the rotations of corresponding joints. The translation and the rotation around the vertical axis are excluded from the root joint, ensuring a normalized body orientation.

4.1.2 Text Representation. Text transcription is an important speech modality that provides high-level linguistic information in a compact format. It is typically given as a word sequence, where the number of words per unit time can vary depending on the speed of the speech. Following [Kucherenko et al. 2020], we align the words to the speech and convert the text into frame-level features to overcome this issue, which is done using an off-the-shelf text-speech alignment tool combined with a pre-trained language model.

Text-speech alignment is a standard technique in the field of speech synthesis. In our system, we employ Montreal Forced Aligner (MFA) [McAuliffe et al. 2017] for this task, which pinpoints the beginning and end frames of every word in the speech. MFA also identifies silences and represents them as empty words. Since a speaker typically stops gesticulating in a long silence [Graziano and Gullberg 2018], our system records those silences and uses them during training to reproduce such behaviors, as will be detailed later.

We then pass the text and the empty words into BERT [Devlin et al. 2019], a popular and powerful pre-trained language model, to extract a high-level representation of the text. BERT computes an encoding vector for each word in an input sentence, which is then repeated and used for all the frames that the word occupies. We represent these word vectors collectively as $\mathbf{T} = \{t_1, \dots, t_K\}$ for a speech clip of K frames, where each $t \in \mathbb{R}^{d_t}$.

4.1.3 Audio Representation. Many recent studies use deep encoders to extract audio features from raw audio signals or audio spectrograms, where only the features extracted at the last layer of the encoder are used to generate gestures [Alexanderson et al. 2020; Kucherenko et al. 2020; Li et al. 2021a; Qian et al. 2021; Yoon et al. 2020]. Such a configuration potentially encourages the encoder to

mix information from multiple levels into the same feature, which can be difficult to disentangle in the downstream generation tasks.

In our system, we propose to decouple the multi-level audio features in the encoder and use them in different scenarios. We assume the high-level features correspond to the speech semantics that determines the gesture lexemes, while the low-level features relate to the other audio information and can be used to control the gesture styles. As shown in Figure 4, we employ a pre-trained speech model, vq-wav2vec [Baevski et al. 2020], to extract audio features from raw audio signals and fine-tune it using a contrastive learning strategy.

The encoder of vq-wav2vec has $L = 8$ convolutional layers. When taking a block of audio signals of K frames, A , as input, this encoder produces a representation \mathbf{a}_k for each frame k of the audio. In this computation, the outputs of every layer can be considered as a set of multi-level features $\{\mathbf{a}_k^l\}, l = 1, \dots, L$, and notably, $\mathbf{a}_k^L = \mathbf{a}_k$. We then encourage the highest-level feature \mathbf{a}_k^L to match the speech content and push apart the features of the lower levels $\{\mathbf{a}_k^l\}, l < L$ to capture crucial content-irrelevant information. Specifically, we utilize the contrastive loss

$$\mathcal{L}_{\text{cont}} = -\log \frac{\exp(\text{sim}(\tilde{\mathbf{t}}_k, \tilde{\mathbf{a}}_k^L)/\tau)}{\sum_{i=1}^K \sum_{l=1}^L \exp(\text{sim}(\tilde{\mathbf{t}}_k, \tilde{\mathbf{a}}_i^l)/\tau)}, \quad (1)$$

where the text feature \mathbf{t}_k is extracted from the speech transcription, the $\text{sim}(\cdot, \cdot)$ function computes the cosine similarity between two vectors as

$$\text{sim}(\tilde{\mathbf{t}}, \tilde{\mathbf{a}}) = \frac{\tilde{\mathbf{t}} \cdot \tilde{\mathbf{a}}}{\|\tilde{\mathbf{t}}\| \|\tilde{\mathbf{a}}\|}, \quad (2)$$

and τ is the temperature hyperparameter. All the feature vectors are projected into the same vector space using learnable linear projections $\tilde{\mathbf{t}} = F_t(\mathbf{t})$ and $\tilde{\mathbf{a}}_k^l = F_a^l(\mathbf{a}_k^l), l = 1, \dots, L$, respectively. Notably, we consider the highest-level audio feature of the current frame as the positive example and audio features of the other levels and the other frames as the negative examples in this contrastive learning process.

This contrastive learning strategy is partially inspired by the HA2G model proposed by Liu et al. [2022]. However, unlike their approach, which considers contrastive learning as a part of the training of the gesture generator, we train the audio encoder in a separate pre-training stage using only the speech data. After the training, the features extracted at the second and the last layers of the encoder, represented by $\mathbf{a}^{\text{low}} \in \mathbb{R}^{d_A}$ and $\mathbf{a}^{\text{high}} \in \mathbb{R}^{d_A}$, respectively, are then used in different training and inference stages in the downstream generation task. They can be represented collectively as \mathbf{A}^{low} and \mathbf{A}^{high} for a speech clip. Although the gesture motions are not considered here, we find that the results of this encoder still demonstrate correlations between the high-level audio features and the gestures. We will discuss these results later in Section 7.4.

4.1.4 Identity Representation. Similar to previous studies [Bhattacharya et al. 2021a; Yoon et al. 2020], our system can leverage the speaker identity (ID) to help distinguish different gesture styles and achieve stylized gesture generation. We represent each speaker as a one-hot vector $\mathbf{I} \in \{0, 1\}^{N_I}$, where N_I is the number of speakers in a dataset.

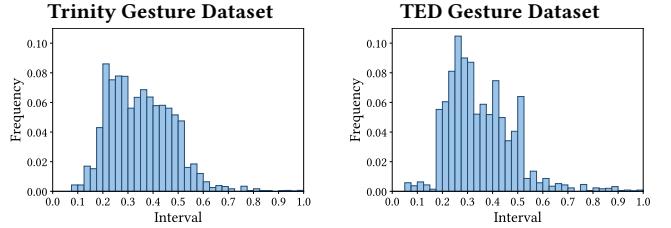


Figure 5: Distribution of time intervals between consecutive audio onsets in two open-source speech-gesture datasets.

4.2 Rhythm-Based Speech Segmentation

In this section, we describe how our system segments and normalizes an input speech into uniform blocks. This procedure is crucial for generating a gesture motion that is temporally synchronized with the rhythm of the speech. To that end, our system first identifies beats in the input audio, which generally corresponds to phonetic properties such as stress or accent, then segments the speech at every beat and time-scales the audio to the same length.

4.2.1 Beat Identification. Rhythm can be characterized by a pattern of beats. In music-related tasks, such as dance generation [Aristidou et al. 2022; Chen et al. 2021], identifying beats using the onsets of audio signals is a standard technique [Bello et al. 2005; Ellis 2007], where off-the-shelf tools such as librosa library [McFee et al. 2015] can be employed to extract those audio features.

However, unlike the rhythm in music that is typically consistent over time, the pattern of beats in a speech can vary significantly according to the context and pace of the speech. Taking a close look at the time intervals between consecutive audio onsets in our training dataset, we notice that the majority of those intervals fall within a range roughly between $D_m = 0.2 \sim 0.3$ seconds and $D_M = 0.5 \sim 0.6$ seconds, as illustrated in Figure 5, though the actual values of D_m and D_M may vary among datasets depending on the personality of the speakers and the language they speak. We also observe that the time intervals shorter than D_m are often caused by noise, filler words, or stuttering. On the other hand, the intervals that are excessively long often correspond to pauses or silent periods.

Based on these observations, our system employs a simple heuristic strategy to identify beats based on the audio onsets. An onset will be recognized as a beat unless the time interval between it and the previous beat is shorter than D_m , in which case the onset will be ignored. If an interval is longer than D_M , we will insert a necessary number of *pseudo-beats* to make the duration of every new interval within the range $[D_m, D_M]$. More specifically, we insert a pseudo-beat at the first frame which is D_m seconds away from any preceding beat and where the volume of the audio is greater than a threshold \tilde{I}_a . Other pseudo-beats are then added recursively in the same way. We set the threshold \tilde{I}_a as the average volume of the environmental noise. If the entire interval is quieter than \tilde{I}_a , a minimal number of pseudo-beats will be placed evenly in it so that each new interval is shorter than D_M .

4.2.2 Normalization. Our system then segments the speech into short clips at every beat. These clips are then time-scaled into uniform blocks of length D_M . The speech modalities are segmented and time-scaled as well in this process. For the motion, M , and text representation, T , of a clip, we resample the corresponding features to match the new length. The audio is processed with additional care, where we use the TSM (Time-Scale Modification) algorithm to change the duration of the audio while preserving the pitch. The audio features A^{low} and A^{high} are then recomputed for the time-scaled audio blocks. The speaker ID I is a constant of the whole speech, which will not be changed during the normalization.

5 GESTURE GENERATION

The generator module is the core component of our system. It synthesizes realistic gesture motions according to a sequence of gesture lexemes, the corresponding style codes, and the low-level features of the audio. In this section, we first introduce how we construct the gesture lexicon and then describe the design and training of the gesture generator.

5.1 Construction of Gesture Lexicon

As revealed in several pieces of literature in linguistics [Kipp 2004; Neff et al. 2008; Webb 1996], only a limited number of lexemes are used in everyday conversation. We assume that each lexeme corresponds to a specific motion category. Our goal is then to extract those motion categories from a large gesture dataset. To achieve this goal, we employ the vector quantized variational autoencoder (VQ-VAE) model [Oord et al. 2017] to learn a categorical representation of the motion and construct the gesture lexicon.

VQ-VAE has been widely used to learn categorical spaces in many successful temporal models [Baevski et al. 2020; Dhariwal et al. 2020; Ramesh et al. 2021; Yan et al. 2021]. Similar to a regular autoencoder, a VQ-VAE also has an encoder-decoder structure but quantizes the latent space using a discrete codebook. The codebook consists of a list of vectors and their associated indices. The output of the encoder network is compared to every vector in the codebook, where the vector that is the closest in Euclidean distance is considered to be the latent representation of the input and will be fed to the decoder. The training of a VQ-VAE is achieved by pulling together the latent code of input and its corresponding codebook vector.

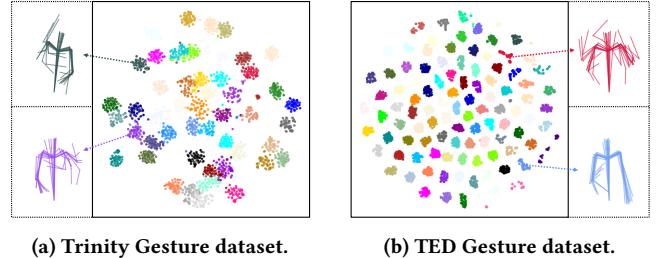
As illustrated in Figure 2, we construct the gesture lexicon by learning the categorical representations of the normalized motion blocks using VQ-VAE. Following [Oord et al. 2017], the loss function is defined as

$$\begin{aligned} \mathcal{L}_{\text{lexicon}} = & \|M - \mathcal{D}(s)\|_2^2 \\ & + w_\alpha \|\mathcal{E}(M) - \text{sg}(s)\|_2^2 + w_\beta \|\text{sg}(\mathcal{E}(M)) - s\|_2^2, \end{aligned} \quad (3)$$

where

$$s = \arg \min_{s' \in \mathcal{S}} \|s' - \mathcal{E}(M)\|_2, \quad (4)$$

M is a normalized motion block, \mathcal{E} and \mathcal{D} represent the encoder and decoder, respectively, sg stands for the *stop gradient* operator that prevents the gradient from backpropagating through it, \mathcal{S} represents the codebook, or the lexicon, and s is a codebook vector, or a lexeme. The first term of Equation (3) penalizes the reconstruction



(a) Trinity Gesture dataset. (b) TED Gesture dataset.

Figure 6: t-SNE visualization of gesture lexicons. Each color stands for a gesture lexeme. (a) lexicon learned on the Trinity Gesture dataset with 50 lexemes. (b) lexicon learned on the TED Gesture dataset with 100 lexemes.

error, while the other two terms pull together the latent representation of motion M and its corresponding lexeme. Notably, since \mathcal{S} is discrete, the $\arg \min$ operator in Equation (4) does not generate a gradient. The gradient of the reconstruction error with respect to the latent code is passed unaltered to the encoder during the backward pass as suggested in [Oord et al. 2017].

We train the VQ-VAE on each dataset in a separate pre-training stage. The encoder is a multi-layer network consisting of four 1-D convolutional layers followed by a fully connected layer, which encodes a motion block into a vector $s \in \mathbb{R}^{d_s}$. The decoder is a mirror of the encoder structurally. The size of the lexicon is a hyperparameter, which is chosen empirically based on the size and complexity of the dataset. Figure 6 shows the t-SNE visualization of the training results on two speech-gesture datasets, along with sample gestures of several lexemes. Once learned, the gesture lexicon and the gesture lexeme of every motion block are fixed and used by the generator and interpreters in both the training and inference stages.

5.2 Architecture of Generator

As illustrated in Figure 2, the generator module of our system is an autoregressive encoder-decoder network, where a new motion block is conditioned on not only the input speech but also the preceding block, the gesture lexeme, and the style code. More specifically, the generation of a motion block M_i can be formulated as

$$M_i^* = \mathcal{G}(\tilde{M}_{i-1}, \tilde{A}_i, s_i, z_i, P), \quad (5)$$

where \tilde{M}_{i-1} represents the features extracted from the preceding motion block, \tilde{A}_i stands for the representation of the input audio, s_i and z_i are the gesture lexeme and style code of the new block, respectively. Note that we use an asterisk (*) to indicate a generated quantity. All these motion and feature blocks have the same length of K frames, where s_i and z_i are repeated and stacked into the corresponding blocks, represented as S_i and Z_i , respectively.

We extract the motion feature $\tilde{M}_{i-1} \in \mathbb{R}^{K \times d_{\tilde{m}}}$ as

$$\tilde{M}_{i-1} = \mathcal{E}_M(M_{i-1}), \quad (6)$$

where the encoder \mathcal{E}_M is a 1-D convolutional network with three layers. The audio feature $\tilde{A}_i \in \mathbb{R}^{K \times d_{\tilde{a}}}$ is computed using three consecutive audio blocks

$$\tilde{A}_i = \mathcal{E}_A(A_{i-1}^{\text{low}}, A_i^{\text{low}}, A_{i+1}^{\text{low}}) \quad (7)$$

to allow the generator to prepare for the future gestures. Notably, the original duration of an audio block is characterized by the onset interval $[D_m, D_M]$, which is typically $[0.2s, 0.5s]$ in our experiments. Thus the temporal window of this encoder is roughly $[0.6s, 1.5s]$. Each A^{low} is the low-level feature pre-computed from the raw speech audio. We assume that A^{low} already captures necessary information and use a simple network consisting of one fully connected layer as the encoder \mathcal{E}_A .

In the training stage, the gesture lexeme s_i of each motion block is determined during the construction of the gesture lexicon, while the style code z_i is a learnable variable that will be trained along with the generator. In the inference stage, both s_i and z_i are provided by the interpreters, as will be discussed below.

In addition to these features, we include a positional encoding block, $P \in \mathbb{R}^{K \times d_P}$, to let the generator know the progress of the generation in a motion block, which is a standard component of transformers [Vaswani et al. 2017] and many sequence generation tasks [Harvey et al. 2020]. For our normalized blocks with K frames, we compute $P = \{p_1, \dots, p_K\}$ as

$$p_{2k} = \sin\left(\frac{K}{\beta^{2k/d_P}}\right) \quad p_{2k+1} = \cos\left(\frac{K}{\beta^{2k/d_P}}\right), \quad (8)$$

where d_P is the dimension of the encoding and $\beta = 10,000$ is a constant controlling the rate of change in frequencies along the embedding dimensions.

The generator \mathcal{G} consists of an MLP-based encoder followed by an LSTM-based decoding network. Inspired by the successful systems in generating sequential output [Oord et al. 2017; Richard et al. 2021], we quantized the latent space of the encoder into H groups of C -way categories, which provides C^H different configurations. We use $H = 64$ and $C = 128$ to ensure a large enough categorical space. In addition, we leverage Gumbel-softmax [Jang et al. 2017] to convert a latent code into a codebook vector, which can be viewed as a differentiable sampling operator for the discrete codebook search.

5.3 Training

We train the generator \mathcal{G} , the encoders \mathcal{E}_M and \mathcal{E}_A , and the learnable style codes $\{z_i\}$ by minimizing a combination of loss terms:

$$\mathcal{L}_{\text{gen}} = w_{\text{rec}} \mathcal{L}_{\text{rec}} + w_{\text{perc}} \mathcal{L}_{\text{perc}} + w_{\text{lexeme}} \mathcal{L}_{\text{lexeme}} + w_z \mathcal{L}_z. \quad (9)$$

The reconstruction loss

$$\mathcal{L}_{\text{rec}} = \|M_i - M_i^*\|_2^2 \quad (10)$$

is simply the MSE loss between the generated motion block and the ground truth. We additionally include a perceptual loss to ensure the similarity between the generated motion and the ground truth in the feature level as well, which is defined as

$$\mathcal{L}_{\text{perc}} = \|\mathcal{E}(M_i) - \mathcal{E}(M_i^*)\|_2^2, \quad (11)$$

where \mathcal{E} is the motion encoder pre-trained in Section 5.1.

We assume that the gesture lexeme determines the type of gesture motion, and the other speech modalities only affect the motion variations. To enforce this assumption, we develop a new perceptual loss, namely the lexeme loss. We first generate a number of new motion blocks using the current gesture lexeme but random

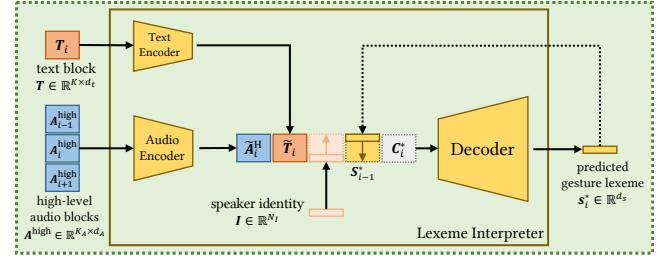


Figure 7: Architecture of the lexeme interpreter.

sets of other features,

$$M_j^* = \mathcal{G}(\tilde{M}_{j-1}, \tilde{A}_j, s_i, z_j, P), \quad (12)$$

where \tilde{M}_{j-1} , \tilde{A}_j , and z_j correspond to a random speech block j . Then the lexeme loss is defined as

$$\mathcal{L}_{\text{lexeme}} = \frac{1}{N_J} \sum_{j \in J} \|s_i - \mathcal{E}(M_j^*)\|_2^2, \quad (13)$$

where J is a random subset of all the motion blocks in the training dataset. The size of J , N_J , is chosen based on the size of the dataset.

Lastly, we regularize the learning of the style code by applying a KL-divergence loss

$$\mathcal{L}_z = D_{KL}(\mathcal{N}(\mu_z, \sigma_z^2) || \mathcal{N}(0, 1)), \quad (14)$$

where μ_z and σ_z^2 are the mean and variance vectors of the style codes in a mini-batch, respectively.

6 CO-SPEECH GESTURE INFERENCE

When given a speech as input, our system segments it into normalized feature blocks $\{A_i^{\text{low}}, A_i^{\text{high}}, T_i, I\}$ and then generates motion blocks $\{M_i^*\}$ recursively, where

$$M_i^* = \mathcal{G}\left(\mathcal{E}_M(M_{i-1}^*), \mathcal{E}_A(A_{i-1}^{\text{low}}, A_i^{\text{low}}, A_{i+1}^{\text{low}}), s_i^*, z_i^*, P\right), \quad (15)$$

and \mathcal{G} , \mathcal{E}_M , \mathcal{E}_A are the components of the learned gesture generator. The generated motion blocks are then denormalized to their original length in the input speech, producing a realistic co-speech gesture animation. Note that we again use the asterisk (*) to indicate a computed quantity that is not provided directly in the speech.

All the variables in Equation (15) are known except the gesture lexeme s^* and style code z^* . As shown in Figure 2, our system learns two interpreters to compute them: the *lexeme interpreter* \mathcal{P}_s translates high-level speech features into the gesture lexemes s^* , and the *style interpreter* \mathcal{P}_z predicts the style code z^* according to the low-level speech features.

6.1 Lexeme Interpreter

As illustrated in Figure 7, the lexeme interpreter is formulated as

$$s_i^* = \mathcal{P}_s(s_{i-1}^*, \tilde{A}_i^H, \tilde{T}_i, I), \quad (16)$$

which is conditioned on the gesture lexeme of the last motion block s_{i-1}^* , and the high-level features of the current speech block. Like the generator, the high-level audio features $\tilde{A}_i^H \in \mathbb{R}^{K \times d_A}$ are computed using three consecutive audio blocks

$$\tilde{A}_i^H = \mathcal{E}_A^{\text{lex}}(A_{i-1}^{\text{high}}, A_i^{\text{high}}, A_{i+1}^{\text{high}}), \quad (17)$$

where each A^{high} contains the high-level representation of the input speech audio, and the encoder \mathcal{E}_A^H is a single-layer fully connected network. The text feature $\tilde{T}_i \in \mathbb{R}^{K \times d_i}$ is also extracted from the text representation of the speech block as

$$\tilde{T}_i = \mathcal{E}_T^{\text{lex}}(T_i), \quad (18)$$

where $\mathcal{E}_T^{\text{lex}}$ is again a single-layer network. Lastly, the one-hot representation of the speaker ID, I , is repeated K times and converted into a feature block.

Those feature blocks can then be concatenated together and fed to an LSTM-based decoder to predict the next gesture lexeme s_i^* . However, considering that the lexemes are selected from the discrete gesture lexicon, we can convert this regression problem into a classification problem. Specifically, instead of directly evaluating Equation (16), we can let \mathcal{P}_s predict the probability that s_i^* is a specific lexeme in the gesture lexicon, then the lexeme with the maximum likelihood will be considered as the result.

6.2 Style Interpreter

The style interpreter shares a similar structure with the lexeme interpreter. It computes z_i^* as

$$z_i^* = \mathcal{P}_z(z_{i-1}^*, s_i^*, \mathcal{E}_T^{\text{style}}(T_i), \mathcal{E}_A^{\text{style}}(A_{i-1}^{\text{low}}, A_i^{\text{low}}, A_{i+1}^{\text{low}})), \quad (19)$$

which is conditioned on the last style code and the new gesture lexeme computed by the lexeme interpreter. The low-level audio representation A^{low} is used in the style interpreter.

6.3 Audio-Only Inference

Both the two interpreters can be reformulated to take only the speech audio as input, where the features related to the text representation T , and optionally the speaker ID I , will be removed from Equation (16) and (19).

In practice, these audio-only interpreters allow cross-language gesture generation, where the speech audio in another language can be taken as input to synthesize realistic gestures without further training. For example, we can utilize a pre-trained model on an English dataset to generate gestures that accompany a Chinese speech. We will show related experiments in Section 7.2.

6.4 Training

During the training of the generator, we have computed the gesture lexeme s_i and the style code z_i of every motion block in the training dataset. We then train the two interpreters using these results as the ground truth. We minimize the standard categorical cross-entropy loss to train the lexeme interpreter, while the MSE loss is used for the style interpreter.

6.4.1 Silent Period Hint. A speaker typically stops gesticulating during a silent pause [Graziano and Gullberg 2018]. Such behaviors are often crucial to the naturalness of a co-speech gesture animation. However, we find that it is often difficult for a gesture generator to deal with silent periods well, even in recent successful systems such as [Alexanderson et al. 2020; Kucherenko et al. 2020]. The speech-gesture datasets may lack necessary motion, and some specific generator models, such as LSTM, may exhibit generative inertia that makes it difficult to become stationary in time.

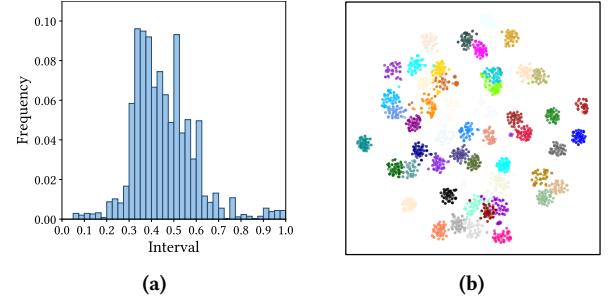


Figure 8: Properties of our Chinese Gesture dataset. (a) Distribution of onset intervals; (b) t-SNE visualization of gesture lexicon.

To solve this problem, we develop a new approach, which we refer to as the *silent period hint*, to encourage the lexeme interpreter to compute a specific *silent lexeme* that corresponds to a silent gesture when encountering a silent period. We check all the lexemes in the lexicon and label a number of stationary ones as the silent lexemes. Notably, the silent lexemes can be automatically labeled by finding such a lexeme corresponding to an empty text word. Then, when a training audio block is in a silent period, which can be detected by the data module of our system, we will force the lexeme interpreter to output the silent lexeme that is the nearest to the current lexeme in the latent space. Moreover, a silent data augmentation is applied when training the generator. We find data blocks that contain empty words and randomly insert 0 ~ 10 consecutive silent blocks after them. The silent block above includes four different features: (a) the audio feature is the environmental noise; (b) the style code is set to zero; (c) the gesture lexeme is the silent lexeme that is the nearest to the previous lexeme in the latent space; and (d) the motion is a stationary pose that is the same as the last frame of the previous motion block. In total, the amount of the inserted silent blocks accounts for 5% of the whole training set.

7 RESULTS

In this section, we first introduce the setup of our system and then evaluate its performance, followed by quantitative and qualitative comparisons with other recent systems. Lastly, we do the ablation study to make an overall analysis of our system.

7.1 System Setup

7.1.1 Datasets. Three speech-gesture datasets are used in this paper: the Trinity dataset [Ferstl and McDonnell 2018], the TED dataset [Yoon et al. 2019], and a Chinese dataset collected for this work.

Trinity Gesture dataset is a large database of speech and gestures jointly collected by Ferstl and McDonnell [2018]. This dataset consists of 242 minutes of motion capture and audio of one male actor talking on different topics. The actor's motion was captured with a 20-camera Vicon system and solved onto a skeleton with 69 joints. In this paper, we use the official release version of *The GENEVA Challenge 2020 Dataset* [Kucherenko et al. 2021a], where

221 minutes are used as training data, and the remaining 21 minutes are kept for testing.

TED Gesture dataset [Yoon et al. 2019] is a 3D upper-body gesture dataset collected using 3D pose estimation from English TED videos. This dataset includes 3D upper-body gestures of the speakers, the aligned speech audios, and text transcripts. In total, there are 253,186 data samples, 80% of which are training samples, 10% belong to the validation set, and the remaining 10% are test samples. The duration of each data sample is 34 frames at a rate of 15 fps, so the total length of this dataset is about 97 hours. Notably, we adapt our model to take 3D joint positions instead of rotations for the TED dataset. The generated gestures are also represented in 3D joint positions, which are then converted into joint rotations for visualization.

Additionally, we collected a 4-hours (80% are the training data and 20% are used for testing) Chinese Gesture dataset using the Noitom Perception Neuron Pro system. This dataset contains 3D full-body gestures of five speakers, aligned speech audios, and Chinese text transcripts. The text transcripts were recognized by Alibaba Cloud Automatic Speech Recognition (ASR) service. The skeleton of this dataset is retargeted to be consistent with the Trinity Gesture dataset. To ensure semantic richness, speakers are instructed to cover a diverse set of topics, such as cooking, fiction, philosophy of life, and academic reporting. Figure 8a illustrates the distribution of onset intervals in our dataset, and Figure 8b shows the visualization of the learned gesture lexicon on our dataset.

7.1.2 System Settings. All the motion data are downsampled to 20, 20, and 15 frames per second on the Trinity, Chinese, and TED datasets, respectively. The range of onset intervals $[D_m, D_M]$ is $[0.2s, 0.5s]$ for both the Trinity and TED datasets, but $[0.3s, 0.6s]$ for the Chinese dataset. The length of each normalized block $K = \lceil D_M \times fps \rceil$. The generator synthesizes 4 seconds of gestures at a time. The dimensions of $d_t, d_{\tilde{a}}, d_{\tilde{A}}, d_s, d_z$, and d_p are 768, 128, 128, 128, 192, 32, and 32 respectively. The size of the gesture lexicon N_s is 50 for both the Trinity and Chinese datasets but 100 for the TED dataset. We train our framework using the Adam optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$ and a learning rate of 0.0003. The loss weights $w_\alpha, w_\beta, w_{rec}, w_{perc}, w_{lexeme}$, and w_z are set as 1.0, 1.0, 1.0, 0.5, 0.2, and 1.0, respectively. At runtime, we use a Gaussian filter with a kernel size of K to smooth the denormalized gesture sequence, where $K = 5$ is chosen to generate the results presented in this paper.

We train separate gesture generators on the Trinity, TED, and Chinese datasets. The cross-language capability of a generator can be further enhanced by pre-training the audio encoder (Section 4.1.3) using datasets in different languages. We have tried pre-training the audio encoder using both an English dataset (such as the Trinity or TED datasets) and our Chinese dataset and using it to train the generator on the Chinese dataset only. The gesture results can be found in the supplementary video.

7.2 Evaluation

Figure 9 shows the gesture synthesis results for the speech excerpts from the test set of the Trinity dataset. Our system generates different types of realistic gestures. The character makes a metaphoric gesture when saying *fine* and an iconic gesture for *defend*. There are

beat gestures for words like *thing* and *selling*, and a deictic gesture appears when the character says *me*.

We also did a cross-language synthesis experiment to test the robustness of our system. We use the pre-trained model trained on the Trinity dataset (an English dataset) to generate gestures for a Chinese speech clip. Since different languages do not share the same word embedding, we generate gestures by taking only the speech audio as input (Section 6.3). As illustrated in Figure 13b, when encountering a different language, our system still generates beat-matched and stylized gestures, reflecting the robustness of our system. We also trained our system on the Chinese dataset and then did another cross-language experiment by generating gestures for an English speech excerpt. Please refer to the supplementary video for the visualization results.

7.2.1 Style Editing. Inspired by the idea of Alexanderson et al. [2020], we augment our system to achieve motion editing by adding a feature block $C \in \mathbb{R}^{K \times d_c}$ of the gesture motion as an extra input of the generator and the lexeme interpreter (Figure 7). The computations in Equation (5) and (16) are then reformulated as

$$M_i^* = \mathcal{G}(\tilde{M}_{i-1}, \tilde{A}_i, s_i, z_i, P, C_i), \quad (20)$$

$$s_i^* = \mathcal{P}_s(s_{i-1}^*, \tilde{A}_i^H, \tilde{T}_i, I, C_i), \quad (21)$$

which allows the network to learn the relationship between a desired motion feature and a gesture motion. During inference phase, we can easily edit the motion style feature of the generated gestures by modifying the feature block C as needed.

Similar to [Alexanderson et al. 2020], we have experimented with three different style features using our system: the height of the right-hand (hand height), the average speed of both hands (hand speed), and the average distance from the hand positions to the up-axis through the root joint of the character (hand radius). We train a separate generator for each style feature. The training data is computed using the reference motions and averaged within a four-second sliding window, forming K -frame feature blocks.

We have synthesized three animations for each of the motion styles. Each animation has a constant desired low, mid, or high feature value, as shown in the first three columns of Figure 10. The last column of Figure 10 shows the accuracy of the generated motion features. These results indicate that all the editing signals could efficiently affect the generated gestures.

7.3 Comparisons

In this section, we compare our system with several state-of-the-art systems to demonstrate the advances made by our system. We first briefly introduce these systems and then our quantitative and qualitative comparisons. We also propose a simple but effective objective metric (PMB) to evaluate the rhythm performance for co-speech gesture synthesis.

7.3.1 Systems of Comparison. We choose four recent successful 3D co-speech gesture synthesis systems for comparison. On the Trinity dataset, we compare with the systems of Style Gesture (SG) [Alexanderson et al. 2020] and Gesticulator (Ges) [Kucherenko et al. 2020]. SG generates gestures based on only the speech audio and uses a normalizing flow model. Ges leverages the audio and text of the speech to generate semantically consistent gestures. On the

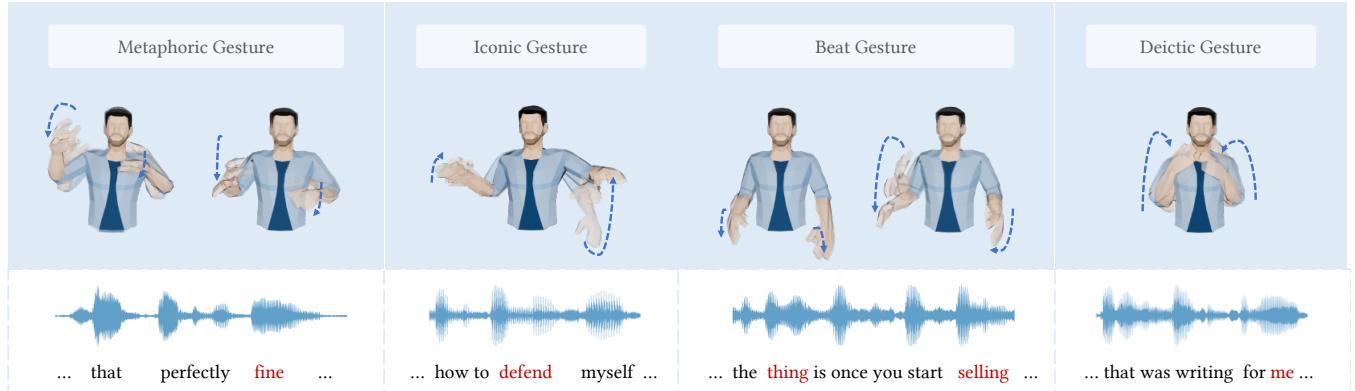


Figure 9: Qualitative results on the gestures synthesized by our method for four sample speech excerpts from the Trinity Gesture dataset [Ferstl and McDonnell 2018]. The character makes a metaphoric gesture when saying *fine* and an iconic gesture for *defend*. There are beat gestures for the words like *thing* and *selling*, and a deictic gesture appears when the character says *me*. The drawing of this figure is inspired by [Yoon et al. 2020].

Table 1: Comparison of our system to SG [Alexanderson et al. 2020], Ges [Kucherenko et al. 2020], GTC [Yoon et al. 2020], and S2AG [Bhattacharya et al. 2021a] on the TED and Trinity datasets. The system without beat segmentation (w/o BC) uses a fixed interval of D_M for segmentation, which is $0.5s \sim 0.6s$ depending on which dataset is used. The system without gesture lexeme (w/o SC) excludes the gesture lexicon and lexeme interpreter modules. The generator is retrained to predict future gestures based on only the previous motion, the audio, and the style code. Similarly, the system without gesture style code (w/o ZC) excludes the style code and the style interpreter modules. Only the motion, the audio, and the lexeme are used by the generator. *Ours (audio only)* denotes the audio-only inference.

Dataset	System	MAJE (mm) \downarrow	MAD (mm/s^2) \downarrow	FGD \downarrow	PMB (%) \uparrow
Trinity	Real Gesture	0.0	0.0	-	95.74
	SG	97.29	4.26	36.98	54.54
	Ges	82.41	3.62	31.04	71.0
	S2AG	54.93	1.49	20.36	79.53
	Ours (w/o BC)	59.11	1.89	16.13	78.18
	Ours (w/o SC)	70.10	2.51	29.75	85.74
	Ours (w/o ZC)	52.85	1.35	12.53	91.36
	Ours (audio only)	57.99	1.83	15.79	87.35
	Ours	49.53	0.97	10.78	91.36
TED	Real Gesture	0.0	0.0	-	93.10
	GTC	26.95	3.03	3.73	71.72
	S2AG	24.49	2.93	3.54	75.57
	Ours (w/o BC)	27.10	3.11	3.88	67.88
	Ours (w/o SC)	30.07	3.53	5.22	83.10
	Ours (w/o ZC)	21.33	2.61	2.47	88.67
	Ours (audio only)	27.28	3.17	3.96	81.33
	Ours	18.13	2.29	2.04	89.52

TED dataset, we compare the systems of Gestures from Trimodal Context (GTC) [Yoon et al. 2020] and Speech to Affective Gestures (S2AG) [Bhattacharya et al. 2021a]. GTC uses speech audio, text transcript, and speaker identity to generate gestures. Based on the three modalities used in GTC, S2AG adds another new modality of affective expressions from the seed poses into their model.

Because we have no access to the official pretrained model of SG, we strictly follow the official configuration and run the codes

offered by authors to train a model. For other systems, we use the pretrained models provided by the authors. For a fair comparison, we use the same skeleton and motion frame rate as the baselines.

7.3.2 Quantitative Evaluation. We first adopt three commonly used evaluation metrics (MAJE, MAD, and FGD) to compare these systems quantitatively. MAJE measures the mean of the absolute errors between the generated joint positions and the ground truth over all

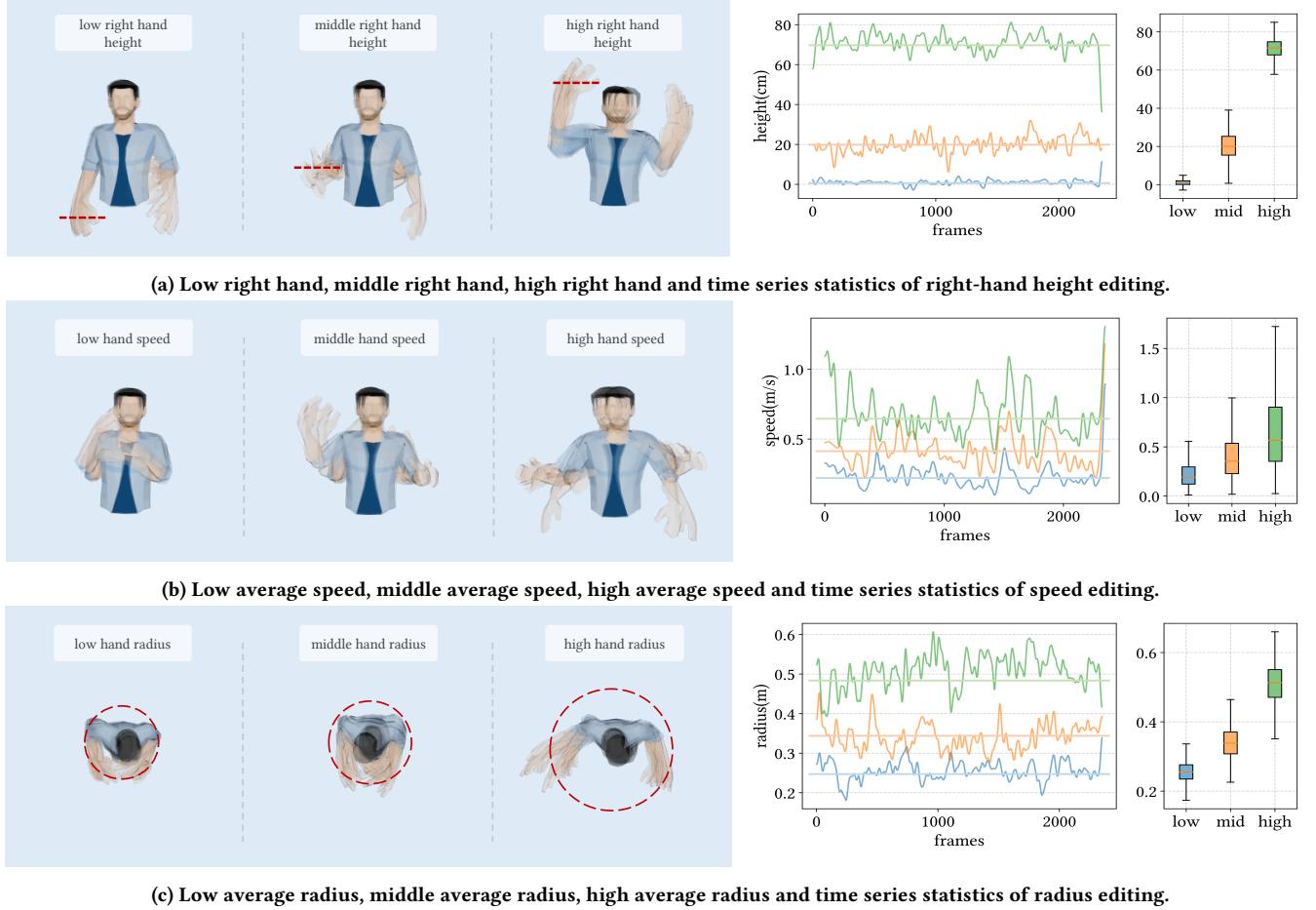


Figure 10: Results of style editing for the right-hand height (the first row), the hand speed (the second row), and the hand radius (the third row). The graphs on the right show the editing input (flat line) and the corresponding values of the output motions. The box plots show the statistics of the output.

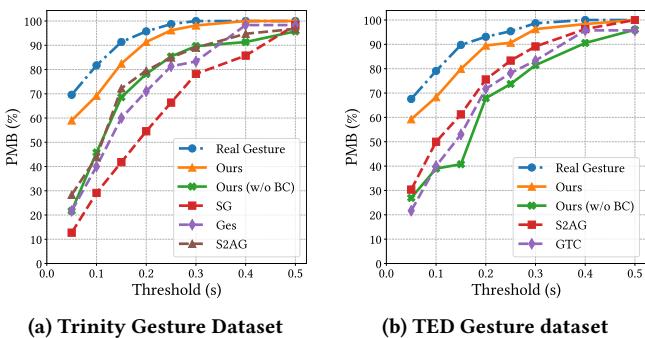


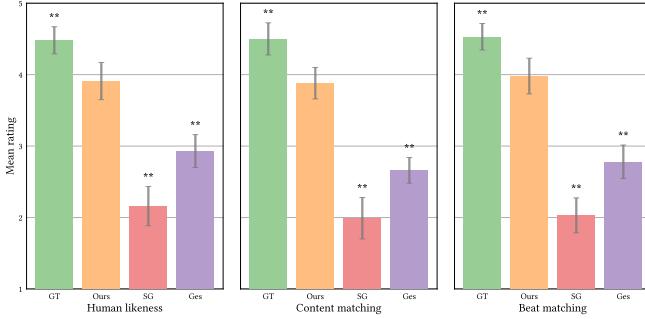
Figure 11: PMB results of continuously adjusting the threshold δ on the Trinity and TED datasets.

the time steps and joints, which indicates how closely the generated joint positions follow the ground truth. MAD measures the mean of the ℓ_2 norm differences between the generated joint accelerations

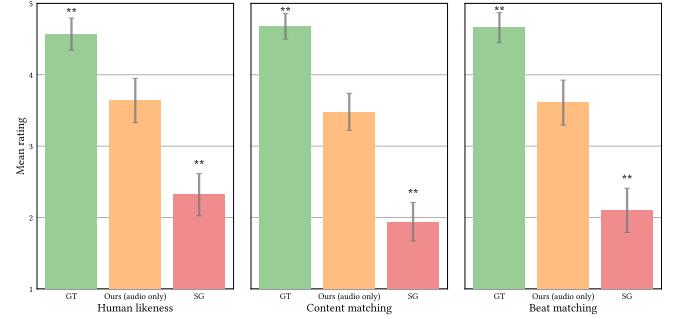
and the ground truth over all the time steps and joints, which indicates how closely the ground truth and the generated joint movements match. Fréchet Gesture Distance (FGD) was proposed by Yoon et al. [2020], which measures the difference between the distributions of the latent features of the generated gestures and ground truth, where the latent features are extracted using an auto-encoder trained on the Human 3.6M dataset [Ionescu et al. 2013]. FGD could assess the perceived plausibility of the synthesized gestures.

Calculating the matching rate between audio and motion beats is a standard method of evaluating rhythm performance, which has been widely used in music-driven dance synthesis [Chen et al. 2021; Li et al. 2021b]. However, we cannot simply apply this method to the gesture generation task because the onset of beat gesture usually precedes the corresponding speech beat by a small amount of time [Pouw and Dixon 2019]. Thus, we need a distance threshold to determine robustly whether the audio and motion beats match each other.

Percentage of correctly predicted keypoints (PCK) is a widely used metric in human pose estimation [Mehta et al. 2017; Wei



(a) Mean ratings for the English speech clips.



(b) Mean ratings for the Chinese speech clips.

Figure 12: User study results with 95% confidence intervals. Asterisks indicate the significant effects (* : $p < 0.05$, ** : $p < 0.01$, * : $p < 0.001$). All the models are trained on the Trinity dataset (an English dataset). See Section 7.3.3 for details.**

et al. 2016], where a predicted key point is considered correct if its distance to the ground truth is smaller than an adjustable threshold. Inspired by this metric, we propose a new metric, PMB , as the percentage of matched beats, where a motion beat is considered to be matched if its temporal distance to a nearby audio beat is smaller than a threshold δ . Specifically,

$$PMB(\mathbf{B}^m, \mathbf{B}^a) = \frac{1}{N_m} \sum_{i=1}^{N_m} \sum_{j=j_{[i-1]}^{i^*}+1}^{N_a} \mathbb{1}[\|\mathbf{b}_i^m - \mathbf{b}_j^a\|_1 < \delta]. \quad (22)$$

The sequence of motion beats $\mathbf{B}^m = \{\mathbf{b}_1^m, \dots, \mathbf{b}_{N_m}^m\}$, where N_m is the number of motion beats, is identified using the algorithm proposed by Ho et al. [2013] based on the local minima of joint deceleration. The audio beats $\mathbf{B}^a = \{\mathbf{b}_1^a, \dots, \mathbf{b}_{N_a}^a\}$ are the onset sequence, where N_a is the number of onsets. $j_{[i-1]}^{i^*}$ represents the index of the audio beats that the last motion beat matches. $\mathbb{1}$ denotes the indicator function. The threshold δ is set to 0.2s by default, but we can adjust δ continuously and observe the changes of the PMB values, which provides a more comprehensive picture of the rhythmic performance.

Table 1 summarizes the performance of all the systems on the two English datasets, Trinity and TED. For the MAJE, MAD, and FGD metrics, our system achieves the lowest values on both datasets. Note that the FGD values of our system are significantly lower than other systems, which indicates the better perceptual quality of gestures synthesized by our system. It is interesting that the FGD values rise rapidly without the gesture lexeme component (w/o SC), which means the gesture lexeme is crucial to gesture quality. There is a decline in the generation performance in the audio-only inference because of the lack of sufficient input information. However, the generated gestures are still acceptable.

As for the rhythm performance, our system achieves the highest PMB values on both datasets. The PMB values drop rapidly without the beat segmentation component (w/o BC), indicating that beat segmentation is vital to rhythm awareness. Figure 11 shows the PMB results with different thresholds. It can be seen that even the ground-truth motion does not match the speech beats precisely, while the behavior of our method is closer to the ground truth compared with the baselines.

7.3.3 User Study. We further conduct user studies to assess the performance of our system qualitatively, where SG [Alexanderson et al. 2020] and Ges [Kucherenko et al. 2020] are used as the baseline methods. We generate 14 test video clips, each consisting of the results synthesized by our methods, the baselines, and the ground truth in random order. Among the 14 clips, nine are English clips generated from the test set of the Trinity dataset, and five are Chinese clips generated from the test set of our Chinese dataset. Notably, all the clips are synthesized using the models trained on the English Trinity dataset. The duration of each clip is around 30s.

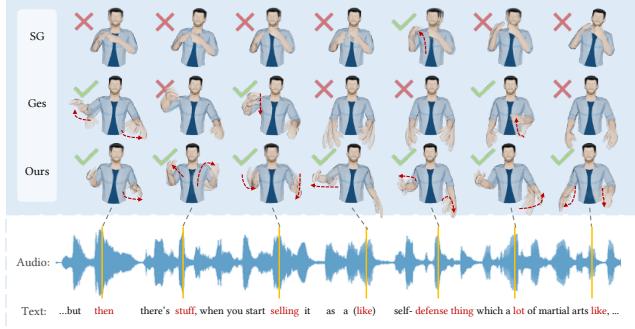
We have recruited 30 volunteer participants to participate in our study, of which 19 are male and 11 are female. 19 participants are 18 – 22 years of age, 10 are between 23 – 30, and one is above 30. When participants watch each video clip, they will be asked to answer three questions and rate the video from 1 to 5, with 1 being the worst, 3 being average, and 5 being the best. The three questions are: (1) human likeness (neglecting the speech audio), (2) speech-to-gesture content matching, and (3) speech-to-gesture beat matching.

The results of these user studies are shown in Figure 12, our system receives higher scores than the other systems and is closer to the real gestures (GT). A one-way ANOVA reveals main effects of *human likeness*, *content matching*, and *beat matching*, and a post-hoc Tukey multiple comparison test identifies a significant difference ($p < 0.005$) between our system and all the other methods. As also illustrated in Figure 13a, the end-to-end systems SG and Ges are less sensitive to rhythm than ours, and the resulting motions lack diversity, which affects their performance in the user studies.

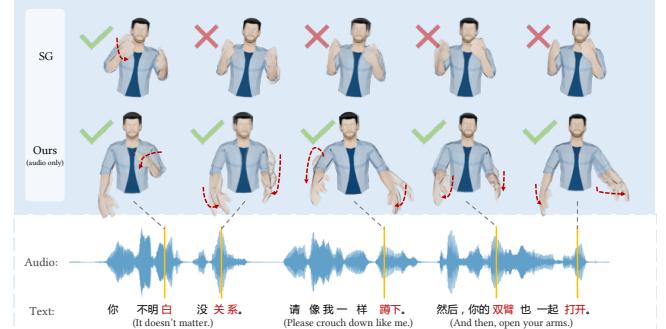
The statistical results of the cross-language test (Figure 12b) demonstrate the better robustness of our system. When dealing with a completely different language (Figure 13b), the gestures generated by SG are more rigid and do not match the beats correctly. In contrast, our model (audio-only) is still able to perceive beats accurately and generate dynamic gestures. Notably, we do not compare with Ges in this cross-language test because this model only supports English text transcripts.

7.4 Ablation Study

We conduct a variety of ablation studies to analyze the performance of our system. Notably, only the ablation of the lexicon size



(a) Gesture results for an English speech clip.



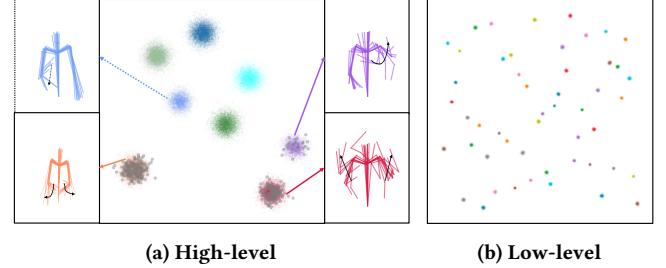
(b) Gesture results for a Chinese speech clip.

Figure 13: Generated motions of SG, Ges, and our system for the same input speech used in the user study. All the models are trained on the Trinity dataset (an English dataset). The red words indicate beats. The red arrows show the movement of corresponding beat gestures. A green check indicates a correct beat match, while a red cross indicates a wrong beat match.

(Figure 18) uses the validation set of the dataset to determine the hyperparameter of the model. All other experiments are based on the test set.

7.4.1 Hierarchical Audio Feature Disentanglement. We presume that the high-level audio feature a^{high} contains semantics-relevant information that determines the gesture lexeme s . To justify this assumption, we apply the K-means algorithm to all the high-level audio blocks of the TED Gesture dataset and get 50 clusters, where each cluster essentially indexes the audio clips with similar semantics. We can find several representative clusters C_0^{high} whose corresponding text transcripts contain words with a similar meaning, such as *many*, *quite a few*, *lots of*, *much*, and *more*, etc. Meanwhile, these audio clips also correspond to a set of generated motion blocks $\{M_0^*, M_1^*, \dots\}$. By encoding these motion blocks using the pre-trained encoder \mathcal{E} , we can obtain their corresponding motion latent codes. As illustrated in Figure 14a, these latent codes (gray dots) only appear in a few gesture lexemes (orange, purple, and red), and it can be seen that the sample gestures of these latent codes indeed convey the semantics of the cluster C_0^{high} . The same observation is not true for the low-level audio features. If we also cluster all low-level audio features and pick a representative cluster C_0^{low} , Figure 14b shows that the corresponding motion latent codes (various color dots) appear nearly uniformly in most of the gesture lexemes. The experiments above confirm the correlation between the high-level audio features and the gesture lexemes, as well as the semantic disentanglement between high-level and low-level audio features.

7.4.2 Gesture Style Code. The low-level audio feature a^{low} contains semantic-irrelevant information, e.g., pitch and volume. Presumably, it should affect the motion variations within a gesture lexeme. In our system, we introduce the learnable gesture style code z combined with the low-level audio feature to jointly determine the motion variations, which can be considered a fine-grained style. As illustrated in Figure 15b, the low-level audio feature increases the variety of the generated gestures but cannot fully decide the variations. With the introduction of the gesture style code (Figure



(a) High-level (b) Low-level

Figure 14: t-SNE visualization of motion latent codes. Each color (except gray) stands for a gesture lexeme. (a) latent codes (gray dots) corresponding to cluster C_0^{high} only appear in specific gesture lexemes (orange, purple, and red). (b) latent codes (various color dots) corresponding to cluster C_0^{low} are distributed in most of the gesture lexemes. See Section 7.4.1 for details.

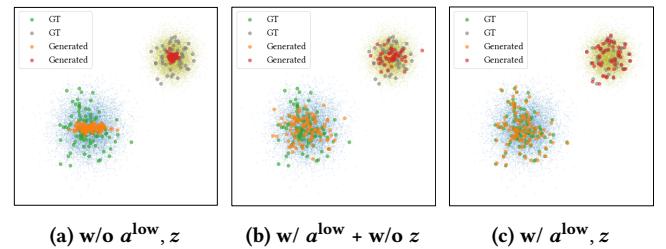


Figure 15: t-SNE visualization of real gestures (GT) versus generated gestures in the motion latent space. We randomly choose two gesture lexemes for visualization. (a) results of our system without low-level audio feature and gesture style code. (b) with low-level audio feature but without gesture style code. (c) with low-level audio feature and gesture style code. See Section 7.4.1 for details.

15c), the distribution of generated gestures becomes closer to the ground-truth distribution.

Table 2: Comparison of style interpreters w/ and w/o low-level audio features and gesture style code. See Section 7.4.1 for details.

Dataset	System	Variance \uparrow	FGD \downarrow	PMB (%) \uparrow
Trinity	Real Gesture	0.41	-	95.74
	w/o a^{low}, z	0.09	17.99	89.35
	w/ $a^{\text{low}} + w/o z$	0.21	12.53	91.36
	w/ $a^{\text{low}}, z^{\text{random}}$	0.30	11.96	91.28
	w/ a^{low}, z	0.37	10.78	91.36
TED	Real Gesture	2.72	-	93.10
	w/o a^{low}, z	0.89	3.71	84.57
	w/ $a^{\text{low}} + w/o z$	1.87	2.47	88.67
	w/ $a^{\text{low}}, z^{\text{random}}$	2.49	2.13	88.89
	w/ a^{low}, z	2.45	2.04	89.52

Table 3: Comparison of style interpreters w/ and w/o text features.

Dataset	System	FGD \downarrow	PMB (%) \uparrow
Trinity	w/o text feature	10.91	91.36
	w/ text feature	10.78	91.36
TED	w/o text feature	2.09	89.22
	w/ text feature	2.04	89.52

Table 2 compares several settings in terms of motion variance, FGD, and PMB. We measure the Euclidean distance from the motion latent code of a motion to the corresponding gesture lexeme and compute the variance of these distances corresponding to the same lexeme. The *motion variance* is then defined as the average of such variances of every lexeme. z denotes the output of the style interpreter, while z^{random} is a random style code sampled from the normal distribution in the latent space. Consistent with Figure 15, Table 2 also shows that combining the low-level audio feature and styles achieves more significant motion variance and lower FGD values. Besides, the FGD values in Table 2 also indicate that the style codes computed by the interpreter generate gestures that are more perceptually acceptable than that created using random style codes.

We have also evaluated the importance of the text features in interpreting the gesture style codes (see Equation 19). The result of Table 3 shows that interpreting the style codes with text features does improve the FGD value. However, the improvement is marginal. Considering the inference efficiency, we can interpret the style codes conditional on only the low-level audio features.

7.4.3 Range of Onset Intervals. Choosing a proper range of onset intervals is crucial to achieving quality gestures. Intuitively, the lower bound of this range affects the model’s sensitivity to beat. If the lower bound is too high (such as 0.5s), the interval between the identified beats becomes large, causing the model to respond sluggishly to beats. The upper bound of the onset intervals regularizes the variance of the duration of the speech blocks, which should

Table 4: Effects of the range of onset intervals.

Dataset	Range of Onset Intervals	FGD \downarrow	PMB (%) \uparrow
Trinity	0.5-1.0s	25.45	73.87
	0.2-1.0s	19.16	90.75
	0.2-0.5s	10.78	91.36
TED	0.5-1.0s	3.61	65.34
	0.2-1.0s	2.55	89.10
	0.2-0.5s	2.04	89.52

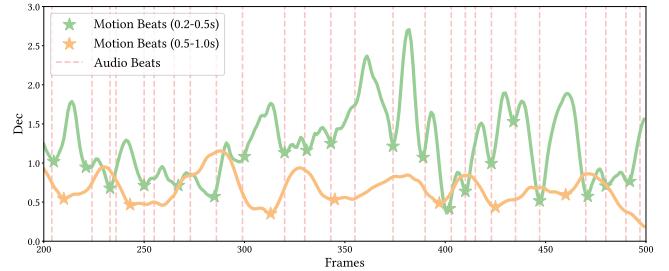


Figure 16: Audio-to-gesture synchronization under different ranges of onset interval. Motion beats are computed based on the local minima of joint deceleration. Audio beats are identified using the audio onsets.

not be too large either. If the length of the speech blocks differs too much, the side effect of the normalization becomes visible, causing unnatural transitions between the generated gesture blocks.

To verify the above hypothesis, we compare the performance of three different interval ranges: 0.2-0.5s, 0.2-1.0s, and 0.5-1.0s, where 0.2-0.5s is our default setting (Section 4.2.1). As shown in Table 4, our default range of onset intervals achieves the best FGD and PMB values on both the Trinity and TED Gesture datasets. PMB drops significantly when the minimum distance between onsets is high (0.5-1.0s), which indicates that the model is insensitive to the rhythm. When the variance is too large (0.2-1.0s), PMB is barely affected, but FGD drops a lot, showing that the generated motions exhibit a lower quality. This result is consistent with our hypothesis.

Moreover, to explicitly show the effect of the minimum onset interval, we visualize the synchronization between the motion and audio beats under different interval ranges in Figure 16. Similar to Aristidou et al. [2022], we calculate the motion beats based on the motion deceleration [Davis and Agrawala 2018]. As shown in Figure 16, the motion beats extracted under the interval range of 0.2-0.5s (denoted by the green stars) synchronize with the audio beats (dashed red lines). In contrast, the motion beats extracted under the interval range of 0.5-1.0s (denoted by the orange stars) do not match well with the audio beats, which proves that a high minimum interval (0.5s) will cause a low beat sensitivity of the model.

7.4.4 Gesture Lexeme. Table 1 has shown that FGD increases significantly without the gesture lexeme, indicating the importance of the gesture lexeme in achieving quality motions. Besides, as demonstrated in Figure 17, the variety of the generated gestures is

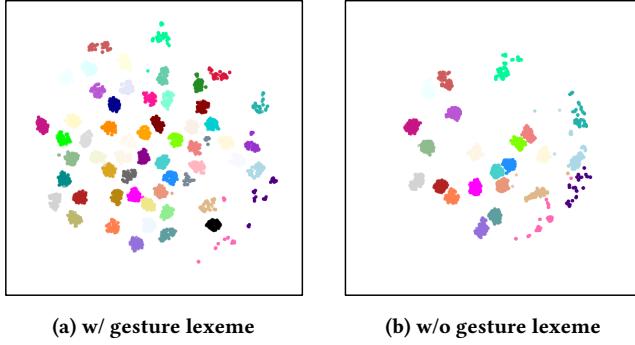


Figure 17: t-SNE visualization of motion latent codes on the TED dataset, computed using the pre-trained encoder \mathcal{E} . Colors represent gesture lexemes.

Table 5: Diversity of motions generated w/ and w/o gesture lexemes.

Dataset	System	Diversity \uparrow
Trinity	Real Gesture	3.79
	w/o gesture lexeme	1.91
	w/ gesture lexeme	3.40
TED	Real Gesture	4.32
	w/o gesture lexeme	2.99
	w/ gesture lexeme	4.09

also significantly reduced without the gesture lexeme. To show this conclusion quantitatively, we calculate the entropy of the gesture lexemes to measure the motion diversity as

$$\text{Diversity} = - \sum_{i=1}^{N_s} p_i \log p_i, \quad (23)$$

where N_s is the size of gesture lexicon, p_i indicates the occurrence frequency of the i -th lexeme in the generated gestures. As shown in Table 5, our system with the gesture lexeme creates much higher diversity than the system without it, which further testifies the conclusion of Figure 17 and again emphasizes the importance of the gesture lexeme.

7.4.5 Size of Gesture Lexicon. Figure 18 shows the performance of our system under different gesture lexicon sizes. It can be seen that neither too small nor too large lexicons achieve good results, measured as FGD values. On the one hand, a small gesture lexicon forces a diverse range of gesture motions to be merged into the same gesture lexeme, which aggravates the one-to-many mapping issue and causes the generator hard to learn all the motions. On the other hand, an excessively large lexicon forcibly splits many lexemes into sub-lexemes. These sub-lexemes are typically close together, making the gesture lexeme interpretation more challenging and thus negatively affecting the gesture quality. We note that the PMB metric is less affected by the gesture lexicon size, possibly because our beat-based segmentation and normalization mechanism explicitly enforces the gesture rhythm. Based on these experiments,

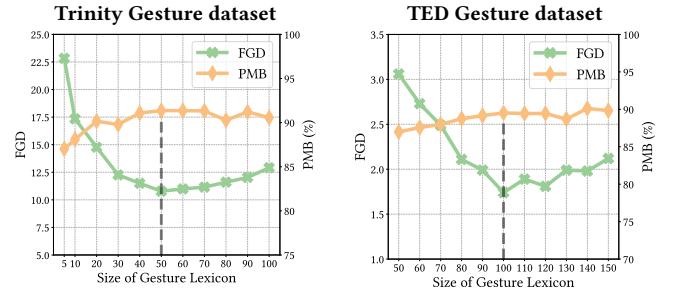


Figure 18: Effects of the size of the gesture lexicon. The gray dashed lines mark the optimal lexicon size.

Table 6: Comparison of interpreters. The statistical interpreter selects gesture lexemes based on the frequency distribution of lexemes but neglects the input speech. The learning-based interpreter is our default interpreter that translates the input speech into gesture lexemes.

Dataset	Interpreter Type	Accuracy (%) \uparrow	FGD \downarrow	PMB (%) \uparrow
Trinity	Statistical	28.01	21.53	90.75
	Learning-based	59.15	10.78	91.36
TED	Statistical	32.11	3.57	87.99
	Learning-based	62.53	2.04	89.52

we set the sizes of the gesture lexicon for the Trinity and TED dataset to 50 and 100, respectively, as shown in Figure 6.

7.4.6 Different Interpreters. In our system, the lexeme interpreter determines the sequence of gesture lexemes. Our learning-based interpreter described in Section 6.1 accomplishes this task according to the speech features and the previous lexemes. Besides, inspired by [Aristidou et al. 2022], we further experiment with a statistical interpreter that matches the frequency of each lexeme in the generated gestures with the reference. Specifically, for a speaker I , we calculate the frequency distribution of gesture lexemes, $f^I \in \mathbb{R}^{N_s}$, and a transition matrix, $L \in \mathbb{R}^{N_s \times N_s}$, describing the frequency of transitions between lexemes using the training data. These quantities can be considered as a global representation of speaker’s gesture style. During inference, we configure the statistical interpreter to ensure that the lexeme distribution of the generated gesture sequence f_t matches the corresponding global distribution f^I . To achieve this goal, at each generation step, we compute a multinomial distribution characterized by f_{t+1}^* using the transition matrix and the difference between the current and target lexeme frequencies f_t and f^I , where

$$f_{t+1}^* = \text{softmax}(f^I - f_t) \cdot L. \quad (24)$$

Then the next lexeme is sampled from the multinomial distribution. This statistical interpreter does not consider the input speech but only the statistics of the generated gestures when selecting the lexemes. In practice, the result motions still look acceptable but

Table 7: Effects of the positional encoding block.

Dataset	System	FGD ↓	PMB (%) ↑
Trinity	w/o positional encoding	11.15	89.98
	w/ positional encoding	10.78	91.36
TED	w/o positional encoding	2.19	88.13
	w/ positional encoding	2.04	89.52

more random. This can be confirmed by Table 6, where the statistical interpreter exhibits a lower prediction accuracy and higher FGD values than our learning-based method because of the lack of semantic information brought by the input speech.

7.4.7 Positional Encoding. The positional encoding block (Equation 8) informs the generator about the frame-level progress of the synthesis in a motion block, which helps the generator model the temporal structure, especially the rhythm, of the sequence. As shown in Table 7, the positional encoding block can improve the beat-matched rate (higher PMB) while enhancing the perceptual quality of generated movements (lower FGD).

8 CONCLUSION

In this paper, we present a rhythm- and semantics-aware co-speech gesture synthesis system that can generate realistic gestures to accompany a speech. For the rhythm, we utilize a segmentation pipeline that explicitly enforces beat alignment to ensure the temporal coherence between the speech and gestures. For the semantics, we successfully disentangle both low- and high-level neural embeddings of speech and motion based on linguistic theory. Then, we devise two neural interpreters to build correspondence between the hierarchical embeddings of the speech and the motion. To evaluate the rhythmic performance, we propose a new objective metric, PMB, to measure the percentage of matched beats. Our method outperforms state-of-the-art systems both objectively and subjectively, as indicated by the MAJE, MAD, FGD, PMB metrics, and human feedback. The cross-language synthesis experiment demonstrates the robustness of our system for rhythmic perception. In terms of application, We show our system’s flexible and effective style editing ability that allows editing of several directorial styles of the generated gestures without manual annotation of the data. Lastly, we have systematically conducted detailed ablation studies that justify the design choices of our system.

There is still room for improvement in our current research. First, our beat detection algorithm is not perfect. We have assumed that the gesture beats coincide with the verbal stresses, but in practice, it has been observed that gesture beats may not always correspond to stressed syllables [McClave 1994]. How to accurately model the complex gestural rhythm is an exciting topic for further exploration. Second, our system can only capture semantics-related gestures repeatedly appearing in the dataset. Learning semantically meaningful gestures that are sparsely distributed in a dataset and allowing a user to control the gesture corresponding to specific semantics is still challenging. Third, our system hypothesizes that each audio onset should correspond to a beat gesture. However, in reality, humans

do not make a beat gesture at every point of verbal emphasis. We believe our framework can be easily augmented by employing another model to predict whether the character should gesture at a specific moment, as suggested by [Speech2Properties2Gestures], and replacing the corresponding lexeme with the silent lexeme. Finally, we only consider the upper body gestures in this work. Generating full-body gestures that include locomotion, facial expressions, finger motions, and the temporal and semantic correspondence among them is a valuable topic for future exploration.

REFERENCES

- Simon Alexanderson, Gustav Eje Henter, Taras Kucherenko, and Jonas Beskow. 2020. Style-Controllable Speech-Driven Gesture Synthesis Using Normalising Flows. *Computer Graphics Forum* 39, 2 (2020), 487–496.
- Andreas Aristidou, Daniel Cohen-Or, Jessica K Hodgins, Yiorgos Chrysanthou, and Ariel Shamir. 2018. Deep motifs and motion signatures. *ACM Transactions on Graphics (TOG)* 37, 6 (2018), 1–13.
- Andreas Aristidou, Anastasios Yiannakidis, Kfir Aberman, Daniel Cohen-Or, Ariel Shamir, and Yiorgos Chrysanthou. 2022. Rhythm is a Dancer: Music-Driven Motion Synthesis with Global Structure. *IEEE Transactions on Visualization and Computer Graphics* (2022), 1–1.
- Alexei Baevski, Steffen Schneider, and Michael Auli. 2020. vq-wav2vec: Self-Supervised Learning of Discrete Speech Representations. In *International Conference on Learning Representations*.
- Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2019. Multimodal Machine Learning: A Survey and Taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41, 2 (2019), 423–443.
- J.P. Bello, L. Daudet, S. Abdallah, C. Duxbury, M. Davies, and M.B. Sandler. 2005. A Tutorial on Onset Detection in Music Signals. *IEEE Transactions on Speech and Audio Processing* 13, 5 (Sept. 2005), 1035–1047.
- Uttaran Bhattacharya, Elizabeth Childs, Nicholas Rewkowski, and Dinesh Manocha. 2021a. Speech2AffectiveGestures: Synthesizing Co-Speech Gestures with Generative Adversarial Affective Expression Learning. In *Proceedings of the 29th ACM International Conference on Multimedia (Virtual Event, China) (MM ’21)*. Association for Computing Machinery, New York, NY, USA, 2027–2036.
- Uttaran Bhattacharya, Nicholas Rewkowski, Abhishek Banerjee, Pooja Guhan, Aniket Bera, and Dinesh Manocha. 2021b. Text2Gestures: A Transformer-Based Network for Generating Emotive Body Gestures for Virtual Agents. *Corr* abs/2101.11101 (2021).
- Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the association for computational linguistics* 5 (2017), 135–146.
- Judee K Burgoon, Thomas Birk, and Michael Pfau. 1990. Nonverbal behaviors, persuasion, and credibility. *Human communication research* 17, 1 (1990), 140–169.
- Justine Cassell, Hannes Högni Vilhjálmsson, and Timothy Bickmore. 2004. Beat: the behavior expression animation toolkit. In *Life-Like Characters*. Springer, 163–185.
- Kang Chen, Zhipeng Tan, Jin Lei, Song-Hai Zhang, Yuan-Chen Guo, Weidong Zhang, and Shi-Min Hu. 2021. Choreomaster: choreography-oriented music-driven dance synthesis. *ACM Transactions on Graphics (TOG)* 40, 4 (2021), 1–13.
- Chung-Cheng Chiu, Louis-Philippe Morency, and Stacy Marsella. 2015. Predicting co-verbal gestures: A deep and temporal modeling approach. In *International Conference on Intelligent Virtual Agents*. Springer, 152–166.
- Abe Davis and Maneesh Agrawala. 2018. Visual rhythm and beat. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 2532–2535.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *north american chapter of the association for computational linguistics* (2019).
- Prafulla Dhariwal, Heewoo Jun, Christine Payne, Jong Wook Kim, Alec Radford, and Ilya Sutskever. 2020. Jukebox: A Generative Model for Music. *ArXiv* abs/2005.00341 (2020).
- Paul Ekman and Wallace V Friesen. 1969. The repertoire of nonverbal behavior: Categories, origins, usage, and coding. *semiotica* 1, 1 (1969), 49–98.
- Daniel PW Ellis. 2007. Beat tracking by dynamic programming. *Journal of New Music Research* 36, 1 (2007), 51–60.
- Ylva Ferstl and Rachel McDonnell. 2018. Investigating the use of recurrent motion modelling for speech gesture generation. In *Proceedings of the 18th International Conference on Intelligent Virtual Agents*. 93–98.
- Shiry Ginosar, Amir Bar, Gefen Kohavi, Caroline Chan, Andrew Owens, and Jitendra Malik. 2019. Learning individual styles of conversational gesture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3497–3506.
- Maria Graziano and Marianne Gullberg. 2018. When Speech Stops, Gesture Stops: Evidence From Developmental and Crosslinguistic Comparisons. *Frontiers in Psychology* 0 (2018).

- David Greenwood, Stephen Laycock, and Iain Matthews. 2017. Predicting head pose from speech with a conditional variational autoencoder. ISCA.
- Ikhсанул Habibie, Weipeng Xu, Dushyant Mehta, Lingjie Liu, Hans-Peter Seidel, Gerard Pons-Moll, Mohamed Elgharib, and Christian Theobalt. 2021. Learning speech-driven 3d conversational gestures from video. In *Proceedings of the 21st ACM International Conference on Intelligent Virtual Agents*. 101–108.
- Félix G. Harvey, Mike Yurick, Derek Nowrouzezahrai, and Christopher Pal. 2020. Robust Motion In-Betweening. *ACM Trans. Graph.* 39, 4, Article 60 (jul 2020), 12 pages.
- Dai Hasegawa, Naoshi Kaneko, Shinichi Shirakawa, Hiroshi Sakuta, and Kazuhiko Sumi. 2018. Evaluation of speech-to-gesture generation using bi-directional LSTM network. In *Proceedings of the 18th International Conference on Intelligent Virtual Agents*. 79–86.
- Gustav Eje Henter, Simon Alexanderson, and Jonas Beskow. 2020. Moglow: Probabilistic and controllable motion synthesis using normalising flows. *ACM Transactions on Graphics (TOG)* 39, 6 (2020), 1–14.
- Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems* 30 (2017).
- Chieh Ho, Wei-Tze Tsai, Keng-Sheng Lin, and Homer H Chen. 2013. Extraction and alignment evaluation of motion beats for street dance. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2429–2433.
- Daniel Holden, Taku Komura, and Jun Saito. 2017. Phase-functioned neural networks for character control. *ACM Transactions on Graphics (TOG)* 36, 4 (2017), 1–13.
- Chien-Ming Huang and Bilge Mutlu. 2012. Robot behavior toolkit: generating effective social behaviors for robots. In *2012 7th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 25–32.
- Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. 2013. Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE transactions on pattern analysis and machine intelligence* 36, 7 (2013), 1325–1339.
- Eric Jang, Shixiang Gu, and Ben Poole. 2017. Categorical Reparameterization with Gumbel-Softmax. *ICLR* (2017).
- Hanbyul Joo, Tomas Simon, Mina Cikara, and Yaser Sheikh. 2019. Towards social artificial intelligence: Nonverbal social signal prediction in a triadic interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10873–10883.
- Adam Kendon. 2004. *Gesture: Visible Action as Utterance*. Cambridge University Press, Cambridge.
- Jae Woo Kim, Hesham Fouad, and James K Hahn. 2006. Making Them Dance.. In *AAAI Fall Symposium: Aurally Informed Performance*, Vol. 2.
- Michael Kipp. 2004. *Gesture Generation by Imitation: From Human Behavior to Computer Character Animation*. Dissertation.com, Boca Raton.
- Stefan Kopp, Brigitte Krenn, Stacy Marsella, Andrew N Marshall, Catherine Pelachaud, Hannes Pirker, Kristinn R Thórisson, and Hannes Vilhjálmsson. 2006. Towards a common framework for multimodal generation: The behavior markup language. In *International workshop on intelligent virtual agents*. Springer, 205–217.
- Lucas Kovar, Michael Gleicher, and Frédéric Pighin. 2002. Motion Graphs. *ACM Transactions on Graphics* 21, 3 (July 2002), 473–482.
- Taras Kucherenko, Dai Hasegawa, Gustav Eje Henter, Naoshi Kaneko, and Hedvig Kjellström. 2019. Analyzing input and output representations for speech-driven gesture generation. In *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents*. 97–104.
- Taras Kucherenko, Patrik Jonell, Sanne van Waveren, Gustav Eje Henter, Simon Alexandersson, Iolanda Leite, and Hedvig Kjellström. 2020. Gesticulator: A framework for semantically-aware speech-driven gesture generation. In *Proceedings of the 2020 International Conference on Multimodal Interaction*. 242–250.
- Taras Kucherenko, Patrik Jonell, Youngwoo Yoon, Pieter Wolfert, and Gustav Eje Henter. 2021a. A Large, Crowdsourced Evaluation of Gesture Generation Systems on Common Data: The GENEVA Challenge 2020. In *26th International Conference on Intelligent User Interfaces (College Station, TX, USA) (IUI '21)*. Association for Computing Machinery, New York, NY, USA, 11–21.
- Taras Kucherenko, Rajmund Nagy, Patrik Jonell, Michael Neff, Hedvig Kjellström, and Gustav Eje Henter. 2021b. Speech2Properties2Gestures: Gesture-Property Prediction as a Tool for Generating Representational Gestures from Speech. In *Proceedings of the 21st ACM International Conference on Intelligent Virtual Agents (Virtual Event, Japan) (IVA '21)*. Association for Computing Machinery, New York, NY, USA, 145–147.
- Sergey Levine, Philipp Krähenbühl, Sebastian Thrun, and Vladlen Koltun. 2010. Gesture Controllers. *ACM Trans. Graph.* 29, 4, Article 124 (jul 2010), 11 pages.
- Jing Li, Di Kang, Wenjie Pei, Xuefei Zhe, Ying Zhang, Zhenyu He, and Linchao Bao. 2021a. Audio2Gestures: Generating Diverse Gestures from Speech Audio with Conditional Variational Autoencoders. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 11293–11302.
- Ruilong Li, Shan Yang, David A. Ross, and Angjoo Kanazawa. 2021b. AI Choreographer: Music Conditioned 3D Dance Generation With AIST++. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 13401–13412.
- Hung Yu Ling, Fabio Zinno, George Cheng, and Michiel van de Panne. 2020. Character Controllers Using Motion VAEs. *ACM Transactions on Graphics* 39, 4 (July 2020), 40:40:1–40:40:12.
- Xian Liu, Qianyi Wu, Hang Zhou, Yinghao Xu, Rui Qian, Xinyi Lin, Xiaowei Zhou, Wayne Wu, Bo Dai, and Bolei Zhou. 2022. Learning Hierarchical Cross-Modal Association for Co-Speech Gesture Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Daniel P. Loehr. 2012. Temporal, Structural, and Pragmatic Synchrony between Intonation and Gesture. *Laboratory Phonology* 3, 1 (May 2012), 71–89.
- Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger. 2017. Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi.. In *Interspeech*, Vol. 2017. 498–502.
- Evelyn McClave. 1994. Gestural beats: The rhythm hypothesis. *Journal of psycholinguistic research* 23, 1 (1994), 45–66.
- Brian McFee, Colin Raffel, Dawen Liang, Daniel P. W. Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. 2015. Librosa: Audio and Music Signal Analysis in Python. *Proceedings of the 14th Python in Science Conference*, 18–24.
- David McNeill. 1992. Hand and Mind. *Advances in Visual Semiotics* (1992), 351.
- Dushyanth Mehta, Srinath Sridhar, Oleksandr Sotnychenko, Helge Rhodin, Mohammad Shafiei, Hans-Peter Seidel, Weipeng Xu, Dan Casas, and Christian Theobalt. 2017. Vnect: Real-time 3d human pose estimation with a single rgb camera. *ACM Transactions on Graphics (TOG)* 36, 4 (2017), 1–14.
- Michael Neff, Michael Kipp, Irene Albrecht, and Hans-Peter Seidel. 2008. Gesture Modeling and Animation Based on a Probabilistic Re-Creation of Speaker Style. *ACM Transactions on Graphics* 27, 1 (March 2008), 5:1–5:24.
- van den Aaron Oord, Oriol Vinyals, and Koray Kavukcuoglu. 2017. Neural Discrete Representation Learning. *ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS 30 (NIPS 2017)* (2017).
- Wim Pouw and James A Dixon. 2019. Quantifying gesture-speech synchrony. In *the 6th gesture and speech in interaction conference*. Universitätsbibliothek Paderborn, 75–80.
- Shenhan Qian, Zhi Tu, Yihao Zhi, Wen Liu, and Shenghua Gao. 2021. Speech Drives Templates: Co-Speech Gesture Synthesis with Learned Templates. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 11077–11086.
- Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-Shot Text-to-Image Generation. In *Proceedings of the 38th International Conference on Machine Learning*. PMLR, 8821–8831.
- Alexander Richard, Michael Zollhoefer, Yandong Wen, de la Fernando Torre, and Yaser Sheikh. 2021. MeshTalk: 3D Face Animation from Speech using Cross-Modality Disentanglement. (2021).
- Najmeh Sadoughi and Carlos Busso. 2018. Novel realizations of speech-driven head movements with generative adversarial networks. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 6169–6173.
- Li Siyao, Weijiang Yu, Tianpei Gu, Chunze Lin, Quan Wang, Chen Qian, Chen Change Loy, and Ziwei Liu. 2022. Bailando: 3D Dance Generation by Actor-Critic GPT With Choreographic Memory. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 11050–11059.
- Robotics Softbank. 2018. Naoqi api documentation. In *2016 IEEE International Conference on Multimedia and Expo (ICME)*, vol. <http://doc.aldebaran.com/2-5/homepepper.html>.
- Sebastian Starke, Ian Mason, and Taku Komura. 2022. DeepPhase: Periodic Autoencoders for Learning Motion Phase Manifolds. *ACM Transactions on Graphics* 41, 4 (July 2022), 136:1–136:13.
- Sebastian Starke, Yiwei Zhao, Taku Komura, and Kazi Zaman. 2020. Local motion phases for learning multi-contact character movements. *ACM Transactions on Graphics (TOG)* 39, 4 (2020), 54–1.
- Guillermo Valle-Pérez, Gustav Eje Henter, Jonas Beskow, Andre Holzapfel, Pierre-Yves Oudeyer, and Simon Alexanderson. 2021. Transflower: probabilistic autoregressive dance generation with multimodal attention. *ACM Transactions on Graphics (TOG)* 40, 6 (2021), 1–14.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. In *Advances in Neural Information Processing Systems*, Vol. 30. Curran Associates, Inc.
- Petra Wagner, Zofia Malisz, and Stefan Kopp. 2014. Gesture and speech in interaction: An overview., 209–232 pages.
- Rebecca A. Webb. 1996. *Linguistic Features of Metaphoric Gestures*. Ph.D. Dissertation. University of Rochester, Rochester, New York.
- Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. 2016. Convolutional Pose Machines. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Pieter Wolfert, Nicole Robinson, and Tony Belpaeme. 2022. A Review of Evaluation Practices of Gesture Generation in Embodied Conversational Agents. *IEEE Transactions on Human-Machine Systems* 52, 3 (2022), 379–389.
- Jing Xu, Wei Zhang, Yalong Bai, Qibin Sun, and Tao Mei. 2022. Freeform Body Motion Generation from Speech. *arXiv preprint arXiv:2203.02291* (2022).
- Wilson Yan, Yunzhi Zhang, Pieter Abbeel, and Aravind Srinivas. 2021. Videogpt: Video generation using vq-vae and transformers. *arXiv preprint arXiv:2104.10157* (2021).

Youngwoo Yoon, Bok Cha, Joo-Haeng Lee, Minsu Jang, Jaeyeon Lee, Jaehong Kim, and Geohyuk Lee. 2020. Speech gesture generation from the trimodal context of text, audio, and speaker identity. *ACM Transactions on Graphics (TOG)* 39, 6 (2020), 1–16.

Youngwoo Yoon, Woo-Ri Ko, Minsu Jang, Jaeyeon Lee, Jaehong Kim, and Geohyuk Lee. 2019. Robots Learn Social Skills: End-to-End Learning of Co-Speech Gesture Generation for Humanoid Robots. In *2019 International Conference on Robotics and Automation (ICRA)*. 4303–4309.