ACT2G: Attention-based Contrastive Learning for Text-to-Gesture Generation

HITOSHI TESHIMA, Kyushu University, Japan
NAOKI WAKE, Microsoft, USA
DIEGO THOMAS, Kyushu University, Japan
YUTA NAKASHIMA, Osaka University, Japan
HIROSHI KAWASAKI, Kyushu University, Japan
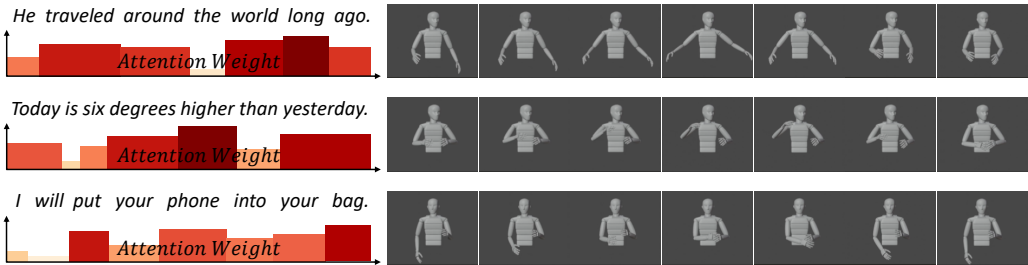KATSUSHI IKEUCHI, Microsoft, USA

Fig. 1. ACT2G takes text as input and outputs realistic gestures. Text is encoded based on the Attention Weight, which represents the likelihood that the gesture will appear, and gesture is generated from the text feature.

Recent increase of remote-work, online meeting and tele-operation task makes people find that gesture for avatars and communication robots is more important than we have thought. It is one of the key factors to achieve smooth and natural communication between humans and AI systems and has been intensively researched. Current gesture generation methods are mostly based on deep neural network using text, audio and other information as the input, however, they generate gestures mainly based on audio, which is called a beat gesture. Although the ratio of the beat gesture is more than 70% of actual human gestures, content based gestures sometimes play an important role to make avatars more realistic and human-like. In this paper, we propose a attention-based contrastive learning for text-to-gesture (ACT2G), where generated gestures represent content of the text by estimating attention weight for each word from the input text. In the method, since text and gesture features calculated by the attention weight are mapped to the same latent space by contrastive learning, once text is given as input, the network outputs a feature vector which can be used to generate gestures related to the content. User study confirmed that the gestures generated by ACT2G were

Authors' addresses: Hitoshi Teshima, teshima.hitoshi.058@s.kyushu-u.ac.jp, Kyushu University, Fukuoka, Japan; Naoki Wake, naoki.wake@microsoft.com, Microsoft, Redmond, Washington, USA; Diego Thomas, thomas@ait.kyushu-u.ac.jp, Kyushu University, Fukuoka, Japan; Yuta Nakashima, n-yuta@ids.osaka-u.ac.jp, Osaka University, Osaka, Japan; Hiroshi Kawasaki, kawasaki@ait.kyushu-u.ac.jp, Kyushu University, Fukuoka, Japan; Katsushi Ikeuchi, katsuike@microsoft.com, Microsoft, Redmond, Washington, USA.

better than existing methods. In addition, it was demonstrated that wide variation of gestures were generated from the same text by changing attention weights by creators.

CCS Concepts: • **Interaction** → *Multimodal Interaction*; *Human-Computer Interfaces*.

Additional Key Words and Phrases: gesture generation, multimodal interaction, contrastive learning

1 INTRODUCTION

In recent years, communication in virtual space has become more active and avatars are increasingly used. In addition tele-operation robots and communication robots become popular and widely developed. Past psychological research has shown that gestures play an important role in conveying information [Birdwhistell 2010; Mcneill 1994], but gesturing avatars and robots is one big challenge. Since manual design of gesture is time-consuming, gesture generation methods have been actively studied for long time. However, it has been extremely difficult to properly reflect the meaning of the content of speech by previous rule based method because the relationship between the semantic information in the speech and the gesture has not been considered.

To solve the problem, learning based approaches have been proposed [Ginosar et al. 2019; Kucherenko et al. 2021a; Li et al. 2021; Qian et al. 2021; Yoon et al. 2019], where gestures are generated from audio or text information trained by using real gesture databases, expecting semantic information being implicitly considered. However, most existing methods generate gestures mainly based on audio, which is called a beat gesture, because the ratio of the beat gesture is more than 70% of actual human gestures [Mcneill 1994]. It should be noticed that text/content based gesture sometimes plays an important role to make gestures more realistic and more human-like.

In the paper, we propose Attention-based Contrastive Learning for Text-to-Gesture (ACT2G), which is a pipeline to generate gestures only from text and explicitly represents the semantic information as shown in Fig. 2. In our technique, to generate the large variation of gestures from arbitrary text, VAE is applied to encode the sequence of gestures into small dimension, and then, texts encoded by a Transformer network are mapped to the same latent space by contrastive learning, by which semantic information is effectively correlated to gestures.

In our method, to generate wide variety of gestures from the same text, but different context, we propose an attention based encoding technique, where the attention weights are estimated from the input word features embedded by BERT and multiplying them by each word feature. The attention weight network is trained by using manually annotated ground truth information of TED Gesture-Type Dataset [Teshima et al. 2022], which is expanded by us and make it publicly available after acceptance. In addition, ACT2G can also generate arbitrary gestures by manually setting attention weight to specific word, where users want to emphasize by the gesture.

The main contributions of our work are the following:

- Contrastive learning for constructing the multi-modal space between gestures and text to achieve semantic gesture generation is proposed
- Attention-based text encoder to focus on specific words which represent gestures is proposed
- Attention-based gesture generating tool based on manual selection of keyword for content creators is developed.
- New gesture database including attention information is created and free for public use.
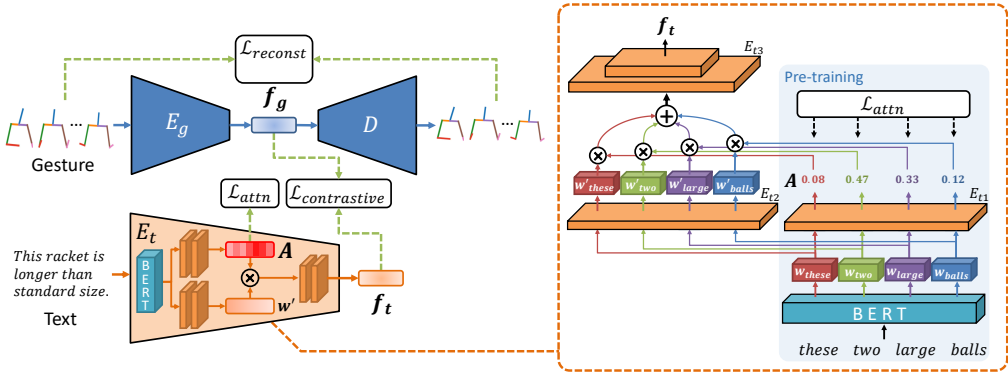
Fig. 2. Pipeline for training. Encoded text feature $\mathbf{f}_t$ and gesture feature $\mathbf{f}_g$ are mapped into a multimodal space by contrastive learning. The figure on the right shows an attention-based text encoder.

2 RELATED WORK

Gestures can be divided into mainly four categories, such as beat, deictic, iconic, and metaphoric [Mcneill 1994]. Beat is a gesture that has nothing to do with the content of the utterance; it is a gesture like shaking arms in time with the inflection of the voice. The other three types of gestures are, respectively, pointing gestures, gestures for concrete objects or actions, and gestures for abstract concepts; these are called representational, which express the content of speech. Absaliev *et al.* denote representational gestures as expressive gestures and analyze the connections between language and expressive gestures. [Abzaliev et al. 2022]. Deep Gesture Generation [Teshima et al. 2022] proposed a method for generating gestures that takes these gesture types into account, however this paper focuses on the generation of representational gestures.

With the remarkable development of deep learning, recent research in gesture generation has tended to be data-driven, with some methods generating gestures from audio, text, or both, or other modalities as well. The trend in gesture generation in recent years has been probabilistic generative models, e.g., adversarial model [Ferstl et al. 2020; Ginosar et al. 2019; Habibie et al. 2021], normalizing flow-based model [Alexanderson et al. 2020], and VAE model [Li et al. 2021]. On the other hand, deterministic models like RNN-based models [Kucherenko et al. 2021a; Takeuchi et al. 2017], Seq2Seq model [Yoon et al. 2019], and auto-encoder model [Lu et al. 2021] also exist. Li *et al.* [Li et al. 2021] pointed out that deterministic generative models to date have been trained with a one-to-one mapping of audio or text to gesture, but because of the diversity of gestures, they trained to separate latent features so that text and gesture were one-to-many. Also, there is a model that predict gesture parameters from speech and refer to appropriate gestures from a database [Ferstl et al. 2021]. Outputting gestures directly from the database enable the generation of gestures that are more human-like. However generating representational gestures from audio alone is difficult because it is hard for the network to learn semantic information related to gestures.

While there are many methods for generating gestures from audio[Ao et al. 2022; Ginosar et al. 2019; Kucherenko et al. 2019; Li et al. 2021; Liu et al. 2022a; Xu et al. 2022], there are also methods for generating gestures from text, such as Seq2Seq [Yoon et al. 2019], transformer model [Bhattacharya et al. 2021], and GPT model [Gao et al. 2023]. A more recent trend is to generate gestures using both text and audio [Ao et al. [n. d.]; Kucherenko et al. 2020; Liang et al. 2022], using the speaker's ID as input [Liu et al. 2022b; Yoon et al. 2020], and also facial expressions and emotions [Liu et al. 2022c]. Ginosar *et al.* focused on generating individual-specific gestures because of the diversity
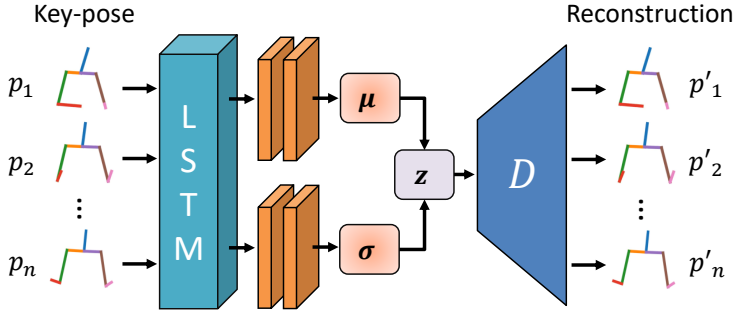
Fig. 3. Gesture-VAE Network. The network takes the key-poses of the gesture as input and predicts the gesture feature **z**.

of gestures [Ginosar et al. 2019], Yoon *et al.* controlled the generated gestures by providing the speaker's ID as input [Yoon et al. 2020]. For those who actually design the gesture, the more input modalities, the higher hurdle to generate gestures. Therefore, we propose a method to generate gestures using only text as input. Our method can also generate a gesture by adding Attention as an input, specifying the words that we want to appear as an representational gesture.

3 PROPOSED METHOD

ACT2G takes text as input, and output a realistic gesture. The training process is divided into three parts: (1) Gesture-VAE, (2) attention-based text encoder, and (3) contrastive learning. The ACT2G pipeline including (2) and (3) is shown in Fig. 2. In Sec. 3.1, we introduce the gesture clustering as a preliminary preparation. Sec. 3.2 introduces the first half of ACT2G, attention-based text encoder, and Sec. 3.3 describes the contrastive learning process for gesture generation.

3.1 Gesture Clustering Using Gesture-VAE

As Li *et al.* mentioned [Li et al. 2021], recent data driven methods take an approach where the input (text, audio, speaker ID, etc.) and the output (gesture) are mapped in a one-to-one fashion. However gestures are diverse, and the same gesture does not always appear in the same text. Inconsistencies in training data can hinder learning. Therefore, in order to achieve a one-to-many correspondences between gesture and text, we apply clustering on gestures in advance.

   We use intermediate features of VAE [Kin 2014] for the clustering of gestures. The network structure is shown in Fig. 3. One of the main problems when clustering time-series data such as gestures is the difference in sequence length. Especially in networks using RNN as encoders, problems such as vanishing gradient make feature extraction of too long data difficult. Therefore, we train the network using the key poses of the gestures as input. Considering key-poses such as labanotation plays an important role in the analysis of human motion, as shown in previous researches [Ikeuchi et al. 2018; Manoj et al. 2009]. We extracted the key-poses $\mathbf{p} = \{\mathbf{p}_1, ..., \mathbf{p}_n\}$ using off-the-shelf algorithm [Ikeuchi et al. 2018] in advance and input them into VAE. The key poses are input to the bi-directional LSTM, where the mean μ and standard deviation σ are estimated and the latent feature $\mathbf{z} \in \mathbb{R}^{32}$ is randomly sampled from the normal distribution corresponding to that parameter. The number of dimensions of $\mathbf{z}$, 32, was determined empirically by plotting ellipses from the mean μ and standard deviation σ and observing the overlap. Then the gesture is reconstructed by decoder D from $\mathbf{z}$. The number of key poses n ranged from 5 to 12 frames, and data were selected from the representational gestures in the TED Gesture-Type Dataset. Each
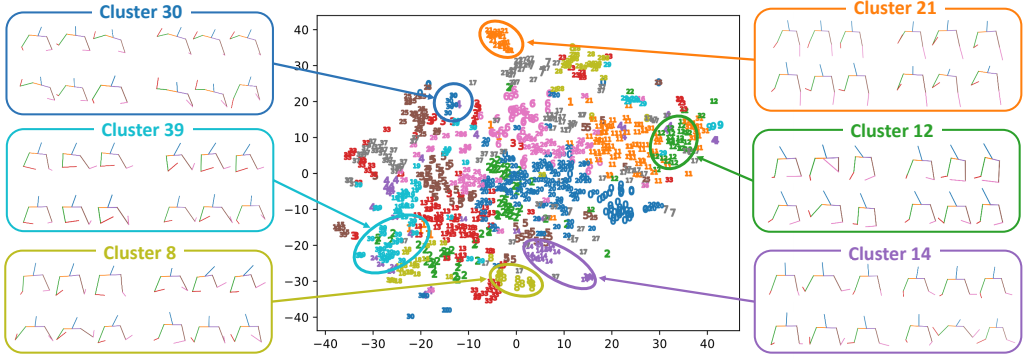
Fig. 4. t-SNE visualization of gesture clustering results. The number of each data point represents the cluster number. Only clusters 30, 39, 8, 21, 12, and 14 show the 3 key-poses of four gestures.

key pose $\mathbf{p}_t$ is represented as relative positions of the 8 upper body joints. The loss function for Gesture-VAE is as follows:

$$\mathcal{L}(\theta, \phi) = \mathbb{E}_{\mathbf{z} \sim q_\theta(\mathbf{z}|\mathbf{p})} \left(\log p\left(\mathbf{p}' \mid \mathbf{z}\right)\right) - D_{KL}\left(q_\theta(\mathbf{z} \mid \mathbf{p}) \| p_\phi(\mathbf{z})\right) \tag{1}$$

where $D_{KL}[.]$ denotes the Kulback-Leibler divergence, $\mathcal{L}$ refers to the likelihood of the parameters of encoder and decoder (i.e., θ and ϕ) and $\mathbf{p}'$ denotes the result of reconstructing the key-poses. Latent feature $\mathbf{z}$ are then used for gesture clustering. Gestures were clustered into 40 clusters by K-Means algorithm.

The results of clustering the gestures are shown in Fig. 4. By extracting gesture features in VAE, each gesture was mapped and clustered in continuous space. For example, cluster 39 and cluster 12 are in opposite positions in Fig. 4, with the gesture of cluster 39 facing left, but the gesture of cluster 12 facing right. In addition, the gesture in cluster 21 has the left hand down, while the gestures in clusters 8 and 14 have the arm up gesture.

Gesture and text in the same cluster were labeled Positive, while data in different clusters were labeled Negative. These positive and negative labels are used during contrastive learning in Sec. 3.3.

3.2 Attention-based Text Encoder

Designing gestures manually requires specialized knowledge and is very labor intensive. Gesture generation using AI simplifies this task. However these approaches cannot represent the gesture as the user would like to design it. Therefore, we propose a method to generate gestures by focusing on words that are likely to be expressed in gestures. Representational gestures in the TED Gesture-Type Dataset are annotated with the corresponding words in the utterance correspond to that gesture. We use this data to estimate attention weight $\mathbf{A}$ that represents the weight of the word to focus on.

Our proposed network structure is shown in the right side of Fig. 2, which is the E_t portion of Fig. 2. First, each word is converted to a word embedding set $\mathbf{w} = \{\mathbf{w}_1, ..., \mathbf{w}_T \mid \mathbf{w}_i \in \mathbb{R}^{768}\}$ by using pre-trained BERT [Devlin et al. 2019]. T is the number of input words, set to 32. Then, the attention weight $\mathbf{A} = \{A_1, ... A_T\}$ is estimated by encoder E_{t1} from $\mathbf{w}$, which consists of a fully connected layer:

$$\mathbf{A} = f\left(E_{t1}\left(\mathbf{w}\right)\right), \tag{2}$$

where f is the following normalization function to ensure that text features are not affected by word count:

$$f_i(x) = \frac{x_i}{\sum_j x_j}. \tag{3}$$

Another word feature $\mathbf{w}'$ is estimated from $\mathbf{w}$ by using encoder E_{t2} and by multiplying with $\mathbf{A}$. During training, $\mathbf{w}$ is always the same value for the same word because BERT is frozen, but $\mathbf{w}'$ is fine-tuned by E_{t2}.

Each word feature is then concatenated and the encoder E_{t3} outputs a text feature $\mathbf{f}_t$ as in the following equations:

$$\mathbf{w}' = E_{t2}(\mathbf{w}) \tag{4}$$

$$\mathbf{f_t} = E_{t3}\left(\|_{i=1}^{T}(\mathbf{A} \odot \mathbf{w}')\right), \tag{5}$$

where $\|_{i=1}^{T}$ represents the vector concatenation from 1 to T and $\odot$ represents the Hadamard product. Attention weight $\mathbf{A}$ is regularized by binary cross entropy loss:

$$\mathcal{L}_{attn} = -\frac{1}{T}\sum_{i=1}^{T}\hat{A}_i \cdot \log A_i + \left(1 - \hat{A}_i\right) \cdot \log\left(1 - A_i\right), \tag{6}$$

where $\hat{A}_i$ is a ground truth with label 1 for the word corresponding to the gesture and 0 for the others. The part on the right side of Fig. 2, illustrate estimation process of $\mathbf{A}$. This part was pre-trained with data from all representational gestures in the TED Gesture-Type Dataset. When training contrastive learning, which is discussed in Sec. 3.3, $\mathbf{A}$ is fine-tuned by using text features $\mathbf{f}_t$ to reconstruct the gesture. During inference, gestures are generated from text features $\mathbf{f}_t$, where $\mathbf{A}$ can be explicitly given. In Sec. 4.5, we describe an experiment in which $\mathbf{A}$ is also input to generate a gesture.

3.3 Contrastive Learning for Multimodal Space Construction

Many recent gesture generation methods output sequences of poses directly from the network, but they are often overly slow or jerky. We assume that the slow movement problem is due to the generator's RNN and autoregressive model. Therefore, we propose a method to create a gesture library and search for the appropriate gesture from the library. ACT2G generates gestures using contrastive learning, which is often used to improve text embedding [Kiros et al. 2014] or image and video retrieval [Bain et al. 2021; Miech et al. 2018].

The overview of the network structure is shown at Fig. 2. Key-poses are extracted from the gesture and input to encoder E_g, which consists of a bi-directional LSTM or FCN. E_g outputs the gesture features $\mathbf{f}_g \in \mathbb{R}^{32}$, and the gesture is reconstructed through the decoder similar to Gesture-VAE, described in Sec. 3.1. Contrastive loss is as follows:

$$\mathcal{L}_{contrastive} = \frac{1}{B}\sum_{B}\left[\frac{1}{2}(\mathbf{P} \odot \mathbf{D})^2 + \frac{1}{2}\max(0, \mathbf{m} - (1 - \mathbf{P}) \odot \mathbf{D})^2\right] \tag{7}$$

where B is the batch size, and $\mathbf{P} \in \mathbb{R}^{B \times B}$ is the positive matrix between each data defined by the gesture clustering described in Sec. 3.1, and is a square matrix of 1 if each data is positive and 0 if negative. $\mathbf{D} \in \mathbb{R}^{B \times B}$ is the L2 distance matrix between text feature $\mathbf{f}_t$ and gesture feature $\mathbf{f}_g$. And $\mathbf{m} \in \mathbb{R}^{B \times B}$ is the margin, which is a square matrix with all elements m. We set $m = 20$ in practice. This contrastive loss means, in the multimodal space, the distance between gesture-text
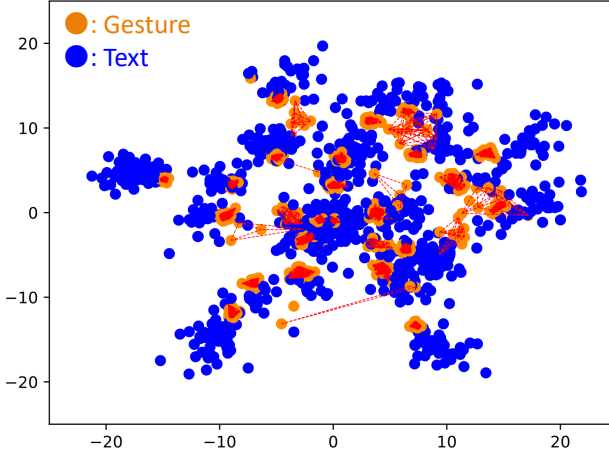
Fig. 5. t-SNE visualization of multimodal space. The blue dots represent text features $\mathbf{f}_t$, the orange dots represent gesture features $\mathbf{f}_g$, and the red dash lines are mutually Positive gesture data pairs.

pairs, defined as positive in Sec. 3.1, is small, while the distance between pairs, defined as negative, is large. The loss to reconstruct the key-poses is:

$$\mathcal{L}_{reconst} = \frac{1}{B} \sum_{i}^{B} \left(p_i' - p_i \right)^2 . \tag{8}$$

The loss function for the entire framework is as follows:

$$\mathcal{L} = \mathcal{L}_{attn} + \alpha \cdot \mathcal{L}_{reconst} + \beta \cdot \mathcal{L}_{contrastive}. \tag{9}$$

Two parameters α and β controls the weights of the loss terms, and they were empirically determined to 10 and 2, respectively. The multimodal space constructed by the contrastive loss is shown in Fig. 5. The blue dots, orange dots, and red lines represent text features $\mathbf{f}_t$, gesture features $\mathbf{f}_g$, and pairs of mutually Positive gestures, respectively. Even though all positive data are connected with each other by a red line, fewer of them are visible as a line compared to the number of dots, indicating that the data that are positive with each other are clustered together.

During inference, using reconstructed key-poses is difficult to humanize the gesture by simply interpolating between the key-poses. Therefore, we propose a method that uses the multimodal space to retrieve appropriate gestures from a gesture library. The gesture library contains the gestures in the training data and their corresponding positions in the multimodal space as shown in Fig. 5. When text is input, the text feature is extracted by the encoder introduced in Sec. 3.2, and the gesture is randomly sampled from nearby that text feature in the multimodal space. Long input text is empirically divided into 8 word and entered into the network. Gesture speed is adjusted to match the length of the human voice, if present, or the synthesized voice, if absent. The gestures generated for each segmented text were combined by spline interpolation.

4   EXPERIMENTS

In this section, we first introduce the dataset we used for training and evaluation in Sec. 4.1. Then, in Sec. 4.2, we describe a ablation study, and in Sec. 4.3, we discuss the evaluation of gestures

generated by our method and state-of-the-art methods. Finally, a gesture generation tool with user specific attention mechanism is demonstrated in Sec. 4.5.

4.1 Dataset

Training ACT2G. The purpose of contrastive learning in ACT2G is to find correlation between texts and gestures and map them into a multimodal space. Therefore, it is necessary to train the network using representational gestures by excluding gestures that are unrelated to text, such as beat gestures. We, therefore used the TED Gesture-Type Dataset [Teshima et al. 2022]. TED Gesture-Type Dataset contains 13,714 gestures divided from TED videos, each annotated with three gesture types: beat, representational, or non-gesture. We used 4097 of these gestures, annotated as representational, for training. When pre-training the attention-based text encoder described in Sec. 3.2, we used all 4097 gestures. And when training the entire ACT2G, we manually annotated most appropriate word representing the gesture for each text of 1000 gestures. For example, an representational gesture which has text of "something that made me very happy" and then pulling the arms back toward one's chest was annotated as "me." Then, we used all the representational gestures for the attention-based text encoder pre-training, and used our original annotated data for the contrastive learning.

Evaluation. Gesture evaluation methods, such as avatars to visualize gestures, question items, and the user interface used for evaluation, vary from method to method. In evaluating gestures, the user study, in which the evaluation is based on human perception, is the most important experiment to focus on, and it is important to evaluate the gestures in the same environment.

Therefore, we used the widely used TED Gesture Dataset [Liu et al. 2022b; Teshima et al. 2022; Yoon et al. 2019] for the user study. We also used the Trinity Speech-Gesture Dataset [Ferstl and Mc-Donnell 2018] for a verification of generalization, according to the GENEA Challenge [Kucherenko et al. 2021b]. The TED Gesture Dataset is data from various people speaking, whereas the Trinity Speech-Gesture Dataset is data of a single speaker speaking on a variety of topics.

The TED Gesture Dataset is used for 2D pose estimation from video by OpenPose [Cao et al. 2019], followed by lifting to 3D [Martinez et al. 2017], or 3D pose estimation by Expose [Choutas et al. 2020]. On the other hand, Trinity Speech-Gesture Dataset uses marker-based motion capture to collect pose information. We divided each dataset into 5 to 15-second sequences, separated by sentence units. Of these, 30 sequences were used for evaluation and 2 sequences were used as attention checks.

4.2 Ablation Study

We first conducted an ablation study to gain more insights into our framework. We used Diversity score [Li et al. 2021] and FGD score [Yoon et al. 2020] as metrics for quantitative evaluation and perceptual scores from the user study for qualitative evaluation. The Diversity proposed by Li *et al.* is not suitable for computation with gesture data of different length, since the distance between gestures is computed with L1 distance of each joint in the same frame. We therefore used the distance between latent features trained in the Gesture-VAE introduced in Section 3.1 as the distance between gestures:

$$\text{Diversity} = \frac{1}{N \times \lceil N/2 \rceil} \sum_{a_1=1}^{N} \sum_{a_2=a_1+1}^{N} \left\| \mu_{a_1} - \mu_{a_2} \right\|_1,$$

where μ refers to the latent vector of the VAE and N is the number of motion clips.

The FGD proposed by Yoon *et al.* [Yoon et al. 2020] could only handle short gestures of 34 frames. Therefore we extracted key poses from the gestures and used them as input to the Gesture-VAE (Fig.

Table 1. Results of ablation study. ± means 95% confidence interval, and ∗ means statistical differences from full model ($p < 0.01$)

	Diversity ↑	FGD ↓	User Score ↑
w/o contrastive	34.99	**155.91**	55.65 ± 1.11∗
w/o attention	36.64	159.85	61.31 ± 1.21∗
full model	**37.43**	163.30	**65.97 ± 1.41**

3), allowing us to evaluate gestures with an average length of 233 frames and a maximum length of 1795 frames. Key poses were extracted using the method of Ikeuchi *et al.* [Ikeuchi et al. 2018] method, which summarizes the entire motion. The gestures, the input to the VAE, were padded to 64 frames and the dimension of the latent space was set to 256 dimensions empirically. During the ablation study, we trained the VAE on the Trinity Gesture Dataset, and tested with 972 motion clips from TED Gesture Dataset.

In the user study, 50 participants in the Amazon Mechanical Turk rated 3 kind of gestures on a scale of 1-100 for the question "How appropriate are the gestures for the speech?". The requirements for participants in Amazon Mechanical Turk and the gestures used in the evaluation are the same as those described in Sec. 4.3 *Comparison with Previous Methods* below.

Table 1 shows the results of the ablation study. W/o contrastive is the case of no contrast learning, i.e., multimodal space is not created. After text features were extracted by BERT, gestures were selected by nearest neighbor method with the features from TED Gesture-Type Dataset. W/o attention means the way the gesture is generated without encoder E_{t1} in the right side of Fig. 2. The result between w/o contrastive and full model in Diversity shows that it is possible to generate a variety of gestures by a multimodal space of text and gestures, rather than a simple nearest neighbor method with only text. The user study results show a statistically significant difference for the full model over the other two ways. The results with w/o contrastive were particularly significant, showing that the multimodal space created by contrastive learning allows for better gesture selection. Although the result of the full model was inferior to the other two methods in terms of the FGD, since FGD just means how similar the gesture is to the original gesture and it is not necessarily an appropriate gesture for speech, we prefer the user score and use the full model in subsequent evaluation.

4.3 Comparison with Previous Methods

We conducted two user studies to compare gestures generated by ACT2G with existing methods and original gestures. The existing methods are compared with (1) Trimodal [Yoon et al. 2020], (2) HA2G [Liu et al. 2022b], (3) Deep Gesture Generation [Teshima et al. 2022] that serve as a baseline, since they are also considering semantic information for gesture generation. (1) **Trimodal** is a method that takes text, audio, and speaker identity as input and generates gestures using bidirectional GRU model. This model is trained with TED Gesture Dataset [Yoon et al. 2019]. (2) **HA2G** is also a method that takes text, audio, and speaker identity as input and generates gestures using decoders that are hierarchically divided into body parts. When inferring, gestures are generated by only the audio and speaker identity as input without text. In both methods Trimodal and HA2G, speaker identities were chosen randomly from the training data, TED Gesture Dataset [Yoon et al. 2019]. (3) **Baseline** takes text as input and uses a gesture library to generate gestures. This model generates gestures by each gesture-type generator after predicting the probability of gesture type (beat, representational, non-gesture) from each word. These methods were evaluated using test data in the TED Gesture Dataset.
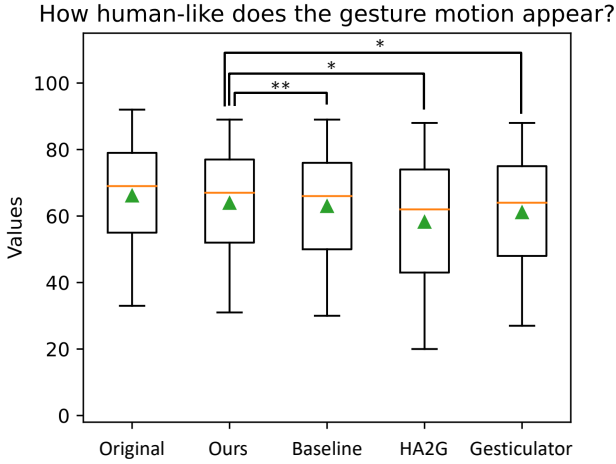
Fig. 6. User study results with Trinity Gesture Dataset. ($*: p < 0.01, **: p < 0.05$)
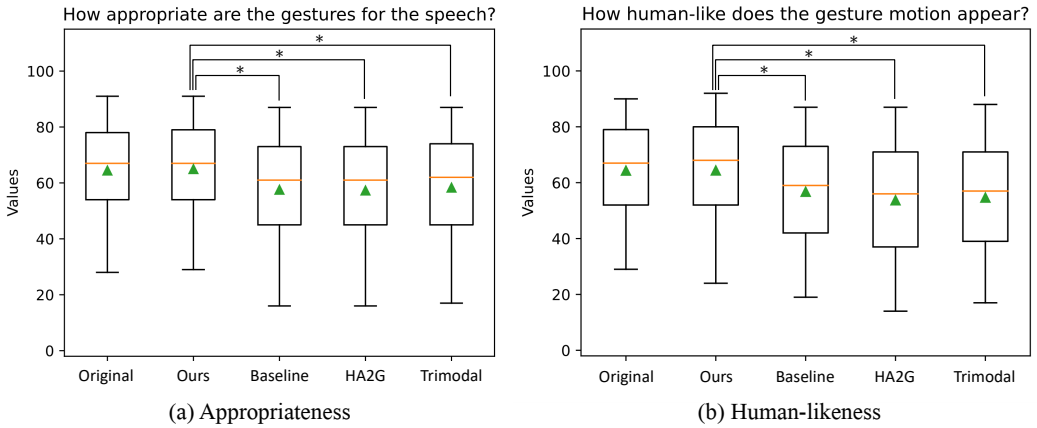

(a) Appropriateness   (b) Human-likeness

Fig. 7. User study results with TED Gesture Dataset. ($*: p < 0.01$) The orange line represents the median, and the green triangle represents the mean. Box edges are at 25 and 75 percentiles, while whiskers cover 95% of all ratings for each.

We built an evaluation environment similar to the GENEA Challenge [Kucherenko et al. 2021b]. As the user interface for evaluating gesture videos, we used HEMVIP [Jonell et al. 2021], which displays multiple videos in parallel and is intuitive and easy to use. Also, the BVH Visualizer [Kucherenko et al. 2021b] was used to visualize the gestures. Since gestures generated by ACT2G, Baseline and HA2G had to be converted to BVH format to use this Visualizer, each joint position was converted to Euler angle. The two questions we prepared for the gesture evaluation items are also the same as in GENEA Challenge, as follows:

(*a*) **Appropriateness** How appropriate are the gestures for the speech?
(*b*) **Human-likeness** How human-like does the gesture motion appear?

When evaluating human-likeness, gestures were evaluated only on the basis of movement, with no audio. Study participants were recruited through the crowdsourcing platform Amazon Mechanical Turk instead of Prolific used in GENEA Challenge. The participants were selected if they satisfied the following three requirements: 1) they had completed more than 500 tasks, 2) their approval rate of task is over 90%, and 3) they had passed the attention check. The attention checks are a way to check the quality of the worker, such as by displaying text such as "Attention! Please rate this video 35" in the some video or replacing the audio. 108 participants met the requirements in the Appropriateness and 115 in the Human-likeness. Participants rated each gesture on a 100-point scale labeled (from best to worst) "Excellent," "Good," "Fair" "Poor" and "Bad" in 20-point intervals. We randomly selected 28 gestures from those annotated as representational gestures in the TED Gesture-Type Dataset with an average original gesture's arm speed above a threshold.

Fig. 7 (a) shows the results of the "Appropriateness", Fig. 7 (b) shows the results of the "Human-likeness". The orange line represents the median value, and the green triangle represents the mean value. The results of ANOVA showed ACT2G was significantly higher than Baseline, HA2G and Trimodal with $p < 0.01$ in both Apprppriateness and Human-likeness indicators. Examples of a generated gesture are shown in Fig. 8. In this figure, the parts where representational gestures might appear are indicated by colored texts and boxes. We can find that Ours has a higher frequency of representational gestures than the other methods. In the example in the upper row, original shows a gesture of spreading his right hand twice, which may represent "chair" or "turn on". On the other hand, our method makes a gesture corresponding to "tie me on the chair" by rotating the left hand, and a gesture corresponding to "loud" by spreading the arms wide. Baseline shows a movement like putting something down, while HA2G and Trimodal show a movement like spreading the right hand and raising both hands to shoulder level when saying "turn on," respectively. In the example in the middle row, original shows the movement of opening both hands when saying "most". Our gesture shows the movement of both arms spread wide when saying "a lot of" and the movement of the right hand to the center when saying "very little". Baseline generates a non-gesture, almost no hand movement. HA2G is gently tightening and opening the elbows, while Trimodal is spreading the left hand down at the "little" part and then opening both arms at the "most" part. In the example in the bottom row, the original gesture is a beat-like gesture with "here", "us", and "all" each with the arms swinging down. In ours, the right hand spreads when saying "right here" and the right elbow rises when saying "all". Baseline and HA2G are both very slow beat gestures. Trimodal is also slow motion and raises the right hand at the "all". The slow movement of previous methods is considered that GRUs and autoregressive models may have caused excessively smooth movements. As for HA2G and Trimodal, there may be some influence of randomly selected speaker identities.

One major issue in gesture generation research is how to establish evaluation criteria. We focused on user scores to evaluate gestures according to the GENEA Challenge [Kucherenko et al. 2021b,c; Yoon et al. 2022], but it is very labor intensive to evaluate each generated gesture with a user study each time. Therefore, many previous studies have quantitatively evaluated gestures, and sought better evaluation metrics. In this study, we report on the better metric in this experiment by comparing how close the previously proposed metrics are to the distribution of user scores. We investigated three metrics used in previous studies: L1 norm [Ginosar et al. 2019; Li et al. 2021], FGD [Liang et al. 2022; Liu et al. 2022b; Yoon et al. 2020], and Jerk [Kucherenko et al. 2019, 2021b]. When calculating L1 norm, the speed of the generated gestures were normalized to make the duration as same as the original gestures. Fig. 9 shows the relationship between user score and each metric, where the user score means the result of the *Appropriateness* described in Sec. 4.3. X-axes of 9(a) and (b) are the L1 norm and the FGD, respectively, and the smaller the better. The correlation coefficients between L1 norm, FGD and user score are -0.0305 and 0.0055, respectively, indicating little correlation. On the other hand, the correlation coefficient for jerk is 0.4459, indicating the
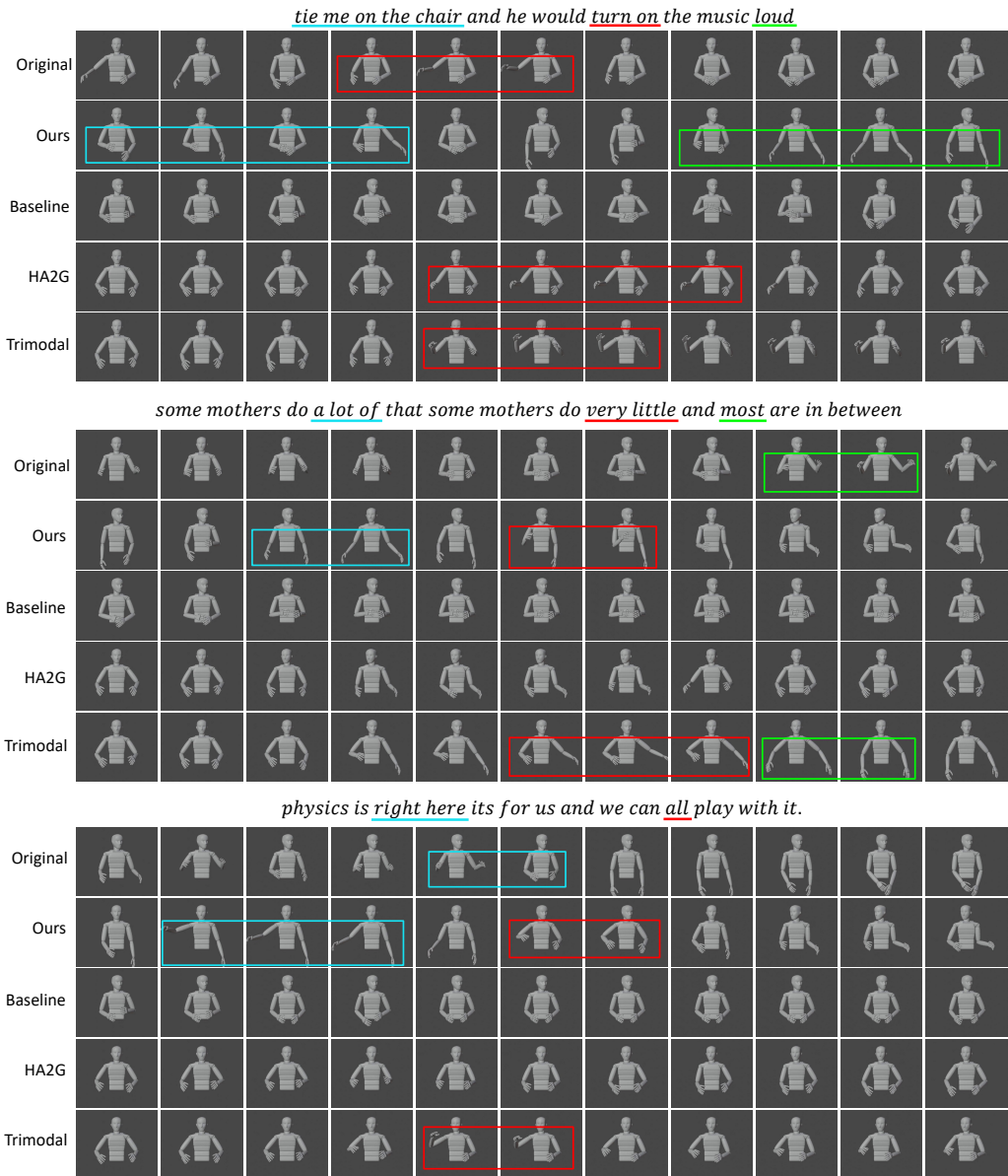
Fig. 8. Qualitative results in TED Gesture Dataset. 3 sequences of input text and 5 kind of gesture pairs. Parts where the representational gestures seem to appear are marked with colored boxes and the corresponding text is also underlined in the same color.

certain correlation between them. We thought that the reason for the trend toward higher ratings with larger jerks might be explained by simple fact that meaningful texts are usually spoken by large movement of human.
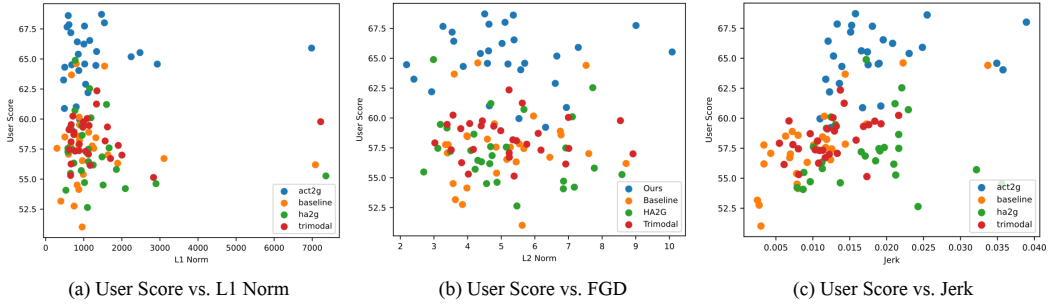
(a) User Score vs. L1 Norm   (b) User Score vs. FGD   (c) User Score vs. Jerk

Fig. 9. Relationship between user score and each metric. (a) User score vs. L1 norm. Correlation coefficient is -0.0305. (b) User score vs. FGD. Correlation coefficient is -0.0055. (c) User score vs. Jerk. Correlation coefficient is 0.4459.

4.4 Verification of Generalization

In Section 4.3 we evaluated four different models trained on the TED Gesture dataset on test data from the same dataset. While the TED Gesture Dataset contains data from a variety of speakers speaking on a variety of topics, the situation in which they are speaking in front of an audience is the same for all data. In addition, the test data did not include anything other than representational gestures such as beat gesture and non-gesture. Therefore, we conducted an additional user study to confirm the generalization performance.

For the test data, we used the Trinity Speech-Gesture Dataset [Ferstl and McDonnell 2018], which is mo-cap data from a single speaker speaking on a variety of topics in the experimental room. The comparison methods include ours, baseline, and HA2G trained with the TED Gesture Dataset, plus Gesticulator [Kucherenko et al. 2020] trained with the Trinity Speech-Gesture Dataset. The experimental environment, including the interface for evaluation and the gesture visualization method, was the same as in the user study in Sec. 4.3, and only human-likeness was used as a question item for participants. We randomly sampled 30 gestures from test data, and 115 participants rate the gestures.

Fig. 6 shows the results for the user study with Trinity Speech-Gesture Dataset. Statistical tests showed Original gesture was significantly best for any other gesture at $p < 0.01$. Ours was significantly superior to Baseline at $p < 0.05$, and to HA2G and Gesticulator at $p < 0.01$. As can be seen from Fig. 9, users tend to prefer the fast-moving (larger movement) gestures when they see five different gestures in parallel. Our method tends to generate fast-moving representational gestures, which is why it was rated higher than the baseline and other methods. Gesticulator outperformed HA2G by p<0.01, possibly because HA2G is a model that is trained on the TED Gesture Dataset, while Gesticulator is a model trained on the Trinity Speech-Gesture Dataset.

4.5 Attention-controlled gesture generation

The attention-based text encoder described in Sec. 3.2 predicts attention, the likelihood that a gesture will appear, for each word. We assumed that by pre-defining the attention corresponding to each word and inputting it into the network, the gesture corresponding to the word with the highest attention weight would appear. Attention weights were set to 0.5 for specified words and 0.1 for other words.

Fig. 10 shows the results with the text "there are a lot of little children there. Fig. 10 (a) shows the result of attention weight **A** estimated by the network with only text input, and (b), (c), and
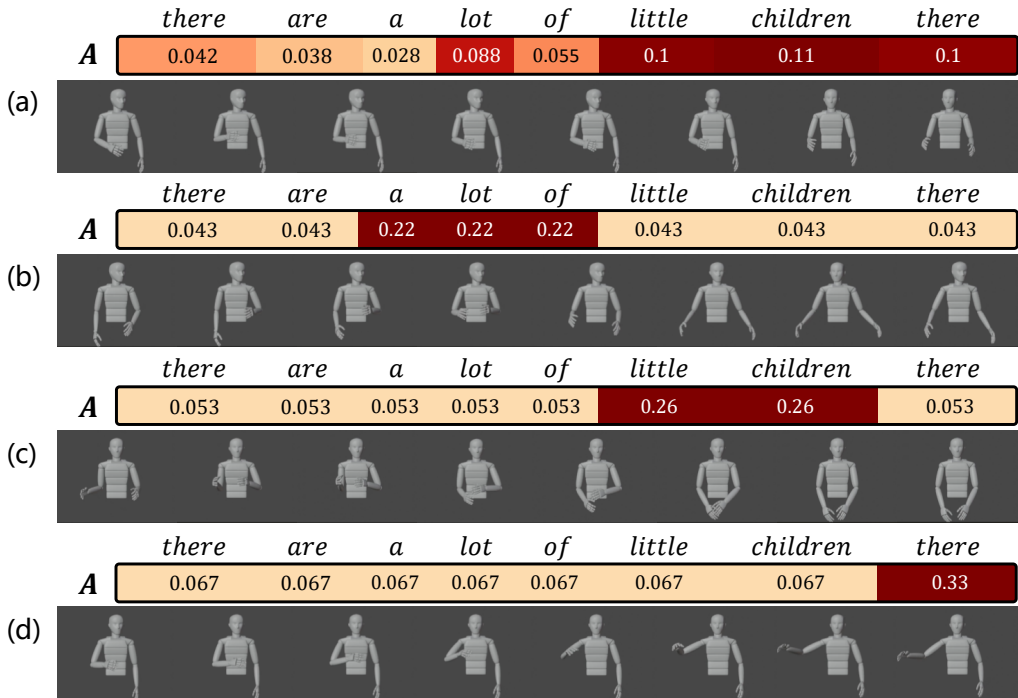
Fig. 10. Qualitative results of gestures generated by inputting attention. (a) Case in which the network predicts **A**. (b) Case with high weight for "a lot of". (c) Case with high weight for "little children". (d) Case with high weight for the second "there".

(d) are the results of generated gestures with higher weights for **A** corresponding to "a lot of", "little children", and second "there", respectively. The each value of **A** is normalized by Equation 3, including the CLS token and padding tokens. In (a) gesture, the right hand is first rotated in a small motion, then both hands are brought forward. This appears to represent a "little" or a second "there". Also, (b), (c), and (d) gestures appear to represent "a lot of" with both arms outstretched, "little children" with arms down, and "there" with right arm extended, respectively. Because of random sampling from the gesture library, it may be generated slightly different gestures even for the same text and attention weight. Although a variety of gestures can be generated by entering text alone, even more diverse gestures can be generated by changing the attention weight.

5 CONCLUSIONS

Gesture generation recently becomes important research topic and many audio based techniques have been proposed, however, few on semantic information based. In this paper, we proposed ACT2G, generating gestures from text, which includes three techniques: (1) gesture clustering based on latent space created by VAE; (2) an attention-based text encoder which explicitly considers words representing representational gestures, and (3) contrastive learning to retrieve content related gestures from the library. In the experiments, user studies were conducted confirming that our method outperformed existing methods in terms of "Appropriateness" and "Human-likeness." Another feature of the attention-based text encoder is that by manually setting the attention weight for each word, it is possible to generate gestures suitable for that word.

Limitations. ACT2G is primarily limited to three aspects. (i) This framework is trained on the TED Gesture-Type Dataset, therefore finger motion is not considered. (ii) Since ACT2G has only been trained on videos of a dozen seconds, a better interpolation method is needed to generate gestures for long sentence input. (iii) The TED Gesture-Type Dataset used for training only contains English videos, therefore gesture-type annotation is needed again for other languages.

ACKNOWLEDGMENTS

REFERENCES

2014. Auto-Encoding Variational Bayes. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings.* arXiv:http://arxiv.org/abs/1312.6114v10 [stat.ML]

Artem Abzaliev, Andrew Owens, and Rada Mihalcea. 2022. Towards Understanding the Relation between Gestures and Language. In *Proceedings of the 29th International Conference on Computational Linguistics.* International Committee on Computational Linguistics, Gyeongju, Republic of Korea, 5507–5520. https://aclanthology.org/2022.coling-1.488

Simon Alexanderson, Gustav Henter, Taras Kucherenko, and Jonas Beskow. 2020. Style-Controllable Speech-Driven Gesture Synthesis Using Normalising Flows. *Computer Graphics Forum* 39 (05 2020), 487–496. https://doi.org/10.1111/cgf.13946

Tenglong Ao, Qingzhe Gao, Yuke Lou, Baoquan Chen, and Libin Liu. 2022. Rhythmic Gesticulator: Rhythm-Aware Co-Speech Gesture Synthesis with Hierarchical Neural Embeddings. 41, 6, Article 209 (nov 2022), 19 pages. https://doi.org/10.1145/3550454.3555435

Tenglong Ao, Zeyi Zhang, and Libin Liu. [n. d.]. GestureDiffuCLIP: Gesture Diffusion Model with CLIP Latents. *ACM Trans. Graph.* ([n. d.]), 18 pages. https://doi.org/10.1145/3592097

Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. 2021. Frozen in Time: A Joint Video and Image Encoder for End-to-End Retrieval. In *IEEE International Conference on Computer Vision.*

Uttaran Bhattacharya, Nicholas Rewkowski, Abhishek Banerjee, Pooja Guhan, Aniket Bera, and Dinesh Manocha. 2021. Text2Gestures: A Transformer-Based Network for Generating Emotive Body Gestures for Virtual Agents**This work has been supported in part by ARO Grants W911NF1910069 and W911NF1910315, and Intel. Code and additional materials available at: https://gamma.umd.edu/t2g. 1–10. https://doi.org/10.1109/VR50410.2021.00037

Ray L. Birdwhistell. 2010. *Kinesics and Context.* University of Pennsylvania Press, Philadelphia. https://doi.org/doi:10.9783/9780812201284

Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. 2019. OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2019).

Vasileios Choutas, Georgios Pavlakos, Timo Bolkart, Dimitrios Tzionas, and Michael J. Black. 2020. Monocular Expressive Body Regression through Body-Driven Attention. In *European Conference on Computer Vision (ECCV).* https://expose.is.tue.mpg.de

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL.*

Ylva Ferstl and Rachel McDonnell. 2018. IVA: Investigating the use of recurrent motion modelling for speech gesture generation. In *IVA '18 Proceedings of the 18th International Conference on Intelligent Virtual Agents.* https://trinityspeechgesture.scss.tcd.ie

Ylva Ferstl, Michael Neff, and Rachel McDonnell. 2020. Adversarial gesture generation with realistic gesture phasing. *Computers and Graphics* 89 (2020), 117–130. https://doi.org/10.1016/j.cag.2020.04.007

Ylva Ferstl, Michael Neff, and Rachel McDonnell. 2021. ExpressGesture: Expressive gesture generation from speech through database matching. *Computer Animation and Virtual Worlds* 32 (2021).

Nan Gao, Zeyu Zhao, Zhi Zeng, Shuwu Zhang, and Dongdong Weng. 2023. GesGPT: Speech Gesture Synthesis With Text Parsing from GPT. *arXiv preprint arXiv:2303.13013* (2023).

S. Ginosar, A. Bar, G. Kohavi, C. Chan, A. Owens, and J. Malik. 2019. Learning Individual Styles of Conversational Gesture. In *Computer Vision and Pattern Recognition (CVPR).* IEEE.

Ikhsanul 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 *ACM International Conference on Intelligent Virtual Agents (IVA).* arXiv:Todo

Katsushi Ikeuchi, Zhaoyuan Ma, Zengqiang Yan, Shunsuke Kudoh, and Minako Nakamura. 2018. Describing Upper-Body Motions Based on Labanotation for Learning-from-Observation Robots. *Int. J. Comput. Vision* 126, 12 (dec 2018), 1415–1429. https://doi.org/10.1007/s11263-018-1123-1

Patrik Jonell, Youngwoo Yoon, Pieter Wolfert, Taras Kucherenko, and Gustav Eje Henter. 2021. HEMVIP: Human Evaluation of Multiple Videos in Parallel. In *Proceedings of the 2021 International Conference on Multimodal Interaction* (Montréal, QC, Canada) *(ICMI '21)*. Association for Computing Machinery, New York, NY, USA, 707–711. https://doi.org/10.1145/3462244.3479957

Ryan Kiros, Ruslan Salakhutdinov, and Richard S. Zemel. 2014. Unifying Visual-Semantic Embeddings with Multimodal Neural Language Models. *ArXiv* abs/1411.2539 (2014).

Taras Kucherenko, Dai Hasegawa, Gustav Eje Henter, Naoshi Kaneko, and Hedvig Kjellström. 2019. Analyzing Input and Output Representations for Speech-Driven Gesture Generation *(IVA '19)*. Association for Computing Machinery, New York, NY, USA, 97–104. https://doi.org/10.1145/3308532.3329472

Taras Kucherenko, Dai Hasegawa, Naoshi Kaneko, Gustav Eje Henter, and Hedvig Kjellström. 2021a. Moving Fast and Slow: Analysis of Representations and Post-Processing in Speech-Driven Automatic Gesture Generation. *International Journal of Human–Computer Interaction* 37, 14 (2021), 1300–1316. https://doi.org/10.1080/10447318.2021.1883883 arXiv:https://doi.org/10.1080/10447318.2021.1883883

Taras Kucherenko, Patrik Jonell, Sanne van Waveren, Gustav Eje Henter, Simon Alexanderson, Iolanda Leite, and Hedvig Kjellström. 2020. Gesticulator: A framework for semantically-aware speech-driven gesture generation. In *Proceedings of the ACM International Conference on Multimodal Interaction*.

Taras Kucherenko, Patrik Jonell, Youngwoo Yoon, Pieter Wolfert, and Gustav Eje Henter. 2021b. A Large, Crowdsourced Evaluation of Gesture Generation Systems on Common Data: The GENEA 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. https://doi.org/10.1145/3397481.3450692

Taras Kucherenko, Patrik Jonell, Youngwoo Yoon, Pieter Wolfert, Zerrin Yumak, and Gustav Henter. 2021c. GENEA Workshop 2021: The 2nd Workshop on Generation and Evaluation of Non-Verbal Behaviour for Embodied Agents. In *Proceedings of the 2021 International Conference on Multimodal Interaction* (Montréal, QC, Canada) *(ICMI '21)*. Association for Computing Machinery, New York, NY, USA, 872–873. https://doi.org/10.1145/3462244.3480983

Jing Li, Di Kang, Wenjie Pei, Xuefei Zhe, Ying Zhang, Zhenyu He, and Linchao Bao. 2021. Audio2Gestures: Generating Diverse Gestures from Speech Audio with Conditional Variational Autoencoders. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)* (2021), 11273–11282.

Yuanzhi Liang, Qianyu Feng, Linchao Zhu, Li Hu, Pan Pan, and Yi Yang. 2022. SEEG: Semantic Energized Co-Speech Gesture Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 10473–10482.

Haiyang Liu, Naoya Iwamoto, Zihao Zhu, Zhengqing Li, You Zhou, Elif Bozkurt, and Bo Zheng. 2022a. DisCo: Disentangled Implicit Content and Rhythm Learning for Diverse Co-Speech Gestures Synthesis. In *Proceedings of the 30th ACM International Conference on Multimedia* (Lisboa, Portugal) *(MM '22)*. Association for Computing Machinery, New York, NY, USA, 3764–3773. https://doi.org/10.1145/3503161.3548400

Haiyang Liu, Zihao Zhu, Naoya Iwamoto, Yichen Peng, Zhengqing Li, You Zhou, Elif Bozkurt, and Bo Zheng. 2022c. BEAT: A Large-Scale Semantic and Emotional Multi-Modal Dataset for Conversational Gestures Synthesis. *arXiv preprint arXiv:2203.05297* (2022).

Xian Liu, Qianyi Wu, Hang Zhou, Yinghao Xu, Rui Qian, Xinyi Lin, Xiaowei Zhou, Wayne Wu, Bo Dai, and Bolei Zhou. 2022b. Learning Hierarchical Cross-Modal Association for Co-Speech Gesture Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10462–10472.

Jinhong Lu, Tianhang Liu, Shuzhuang Xu, and Hiroshi Shimodaira. 2021. Double-DCCCAE: Estimation of Body Gestures From Speech Waveform. *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2021), 900–904.

P. V. Manoj, Kudoh Shunsuke, and Ikeuchi Katsushi. 2009. Keypose and Style Analysis Based on Low-dimensional Representation (Computer Vision and Image Media(CVIM) Vol.2009-CVIM-167).

Julieta Martinez, Rayat Hossain, Javier Romero, and James J. Little. 2017. A simple yet effective baseline for 3d human pose estimation. In *ICCV*.

David Mcneill. 1994. Hand and Mind: What Gestures Reveal About Thought. *Bibliovault OAI Repository, the University of Chicago Press* 27 (06 1994). https://doi.org/10.2307/1576015

Antoine Miech, Ivan Laptev, and Josef Sivic. 2018. Learning a Text-Video Embedding from Incomplete and Heterogeneous Data. (04 2018).

Shenhan Qian, Zhi Tu, YiHao Zhi, Wen Liu, and Shenghua Gao. 2021. Speech Drives Templates: Co-Speech Gesture Synthesis with Learned Templates. *International Conference on Computer Vision (ICCV)*.

Kenta Takeuchi, Dai Hasegawa, Shinichi Shirakawa, Naoshi Kaneko, Hiroshi Sakuta, and Kazuhiko Sumi. 2017. Speech-to-Gesture Generation: A Challenge in Deep Learning Approach with Bi-Directional LSTM. 365–369. https://doi.org/10.1145/3125739.3132594

Hitoshi Teshima, Naoki Wake, Diego Thomas, Yuta Nakashima, Hiroshi Kawasaki, and Katsushi Ikeuchi. 2022. Deep Gesture Generation for Social Robots Using Type-Specific Libraries. In *2022 IEEE/RSJ International Conference on Intelligent Robots*

and Systems (IROS).

Jing Xu, Wei Zhang, Yalong Bai, Qibin Sun, and Tao Mei. 2022. Freeform body motion generation from speech. *arXiv preprint arXiv:2203.02291* (2022).

Youngwoo Yoon, Bok Cha, Joo-Haeng Lee, Minsu Jang, Jaeyeon Lee, Jaehong Kim, and Geehyuk Lee. 2020. Speech Gesture Generation from the Trimodal Context of Text, Audio, and Speaker Identity. *ACM Transactions on Graphics* 39, 6 (2020).

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

Youngwoo Yoon, Pieter Wolfert, Taras Kucherenko, Carla Viegas, Teodor Nikolov, Mihail Tsakov, and Gustav Eje Henter. 2022. The GENEA Challenge 2022: A Large Evaluation of Data-Driven Co-Speech Gesture Generation. In *Proceedings of the 2022 International Conference on Multimodal Interaction* (Bengaluru, India) *(ICMI '22)*. Association for Computing Machinery, New York, NY, USA, 736–747. https://doi.org/10.1145/3536221.3558058