BEAT: A Large-Scale Semantic and Emotional Multi-Modal Dataset for Conversational Gestures Synthesis

Haiyang Liu[1], Zihao Zhu[2], Naoya Iwamoto[3], Yichen Peng[4],
Zhengqing Li[3], You Zhou[3], Elif Bozkurt[5], Bo Zheng[3]

[1]The University of Tokyo. [2]Keio University.
[2]Digital Human Lab, Huawei Technologies Japan K.K.
[4]Japan Advanced Institute of Science and Technology. [5]Huawei Turkey R&D Center.

Fig. 1. Overview. BEAT is a large-scale, multi-modal mo-cap human gestures dataset with semantic, emotional annotations, diverse speakers and multiple languages.

Abstract. Achieving realistic, vivid, and human-like synthesized conversational gestures conditioned on multi-modal data is still an unsolved problem due to the lack of available datasets, models and standard evaluation metrics. To address this, we build **B**ody-**E**xpression-**A**udio-**T**ext dataset, **BEAT**, which has i) 76 hours, high-quality, multi-modal data captured from 30 speakers talking with eight different emotions and in four different languages, ii) 32 millions frame-level emotion and semantic relevance annotations. Our statistical analysis on BEAT demonstrates the correlation of conversational gestures with *facial expressions*, *emotions*, and *semantics*, in addition to the known correlation with *audio*, *text*, and *speaker identity*. Based on this observation, we propose a baseline model, **Ca**scaded **M**otion **N**etwork (**CaMN**), which consists of above six modalities modeled in a cascaded architecture for gesture synthesis. To evaluate the semantic relevancy, we introduce a metric, Semantic Relevance Gesture Recall (**SRGR**). Qualitative and quantitative experiments demonstrate metrics' validness, ground truth data quality, and baseline's state-of-the-art performance. To the best of our knowledge, BEAT is the largest motion capture dataset for investigating human gestures, which may contribute to a number of different research fields, including controllable gesture synthesis, cross-modality analysis, and emotional gesture recognition. The data, code and model are available on https://pantomatrix.github.io/BEAT/.

1   Introduction

Synthesizing conversational gestures can be helpful for animation, entertainment, education and virtual reality applications. To accomplish this, the complex relationship between speech, facial expressions, emotions, speaker identity and semantic meaning of gestures has to be carefully considered in the design of the gesture synthesis models.

While synthesizing conversational gestures based on audio [32,52,20] or text [53,8,5,3] has been widely studied, synthesizing realistic, vivid, human-like conversational gestures is still unsolved and challenging for several reasons. i) **Quality and scale of the dataset**. Previously proposed methods [52,32] were trained on limited mo-cap datasets [46,17] or on pseudo-label [20,52,21] datasets (*cf.* Table 1), which results in limited generalization capability and lack of robustness. ii) **Rich and paired multi-modal data**. Previous works adopted one or two modalities [20,53,52] to synthesize gestures and reported that conversational gestures are determined by multiple modalities together. However, due to the lack of paired multi-modal data, the analysis of other modalities, *e.g.*, facial expression, for gesture synthesis is still missing. iii) **Speaker style disentanglement**. All available datasets, as shown in Table 1, either have only a single speaker [17], or many speakers but different speakers talk about different topics [21,52,20]. Speaker-specific styles were not much investigated in previous studies due to the lack of data. iv) **Emotion annotation**. Existing work [7] analyzes the emotion-conditioned gestures by extracting implicit sentiment features from texts. Due to the unlabeled, limited emotion categories in the dataset [52], it cannot cover enough emotion in daily conversations. v) **Semantic relevance**. Due to the lack of semantic relevance annotation, only a few works [31,52] analyze the correlation between generated gestures and semantics though listing subjective visualization examples. It will enable synthesizing context-related meaningful gestures if existing semantic labels of gestures. In conclusion, the absence of a large-scale, high-quality multi-modal dataset with semantic and emotional annotation is the main obstacle to synthesizing human-like conversational gestures.

There are two design choices for collecting unlabeled multi-modal data, i) the pseudo-label approach [20,52,21], *i.e.*, extracting conversational gestures, facial landmark from in-the-wild videos using 3D pose estimation algorithms [12] and ii) the motion capture approach [17], *i.e.*, recording the data of speakers through predefined themes or texts. In contrast to the pseudo-labeling approach, which allows for low-cost, semi-automated access to large-scale training data, *e.g.*, 97h [52], motion-captured data requires a higher cost and more manual work resulting in smaller dataset sizes, *e.g.*, 4h [17]. However, Due to the motion capture can be strictly controlled and designed in advance, it is able to ensure the quality and diversity of the data, e.g., eight different emotions of the same speaker, and different gestures of 30 speakers talking in the same sentences. Besides, high-quality motion capture data are indispensable to evaluate the effectiveness of pseudo-label training.

Based on the above analysis, to address these data-related problems, we built a mo-cap dataset **BEAT** containing semantic and eight different emotional an-

Table 1. Comparison of Datasets. We compare with all 3D conversational gesture and face datasets. "#", "LM" and "BSW" indicate the number, landmark and blendshape weight, respectively. best and second are highlighted. Our dataset is the largest mocap dataset with multi-modal data and annotations

dataset	Quailty	Modality						Annotation		Scale	
		#body	#hand	face	audio	text	#speaker	#emo	sem	#seq	dura
TED [52]	pseudo	9	-	-	En	✓	>100	-	-	1400	97h
S2G [20,21]	label	14	42	2D LM	En	-	6	-	-	N/A	33h
MPI [47]		23	-	-	-	✓	1	11	-	1408	1.5h
VOCA [16]		-	-	3D Mesh	En	-	12	-	-	480	0.5h
Takechi [46]	mo-cap	24	38	-	Jp	-	2	-	-	1049	5h
Trinity [17]		24	38	-	En	✓	1	-	-	23	4h
BEAT (Ours)	mo-cap	27	48	3D BSW	E/C/S/J	✓	30	8	✓	2508	76h

notations (*cf.* Figure 1), from 30 speakers in four modalities of **B**ody-**E**xpression-**A**udio-**T**ext, annotated in total of 30M frames. The motion capture environment is strictly controlled to ensure quality and diversity, with 76 hours and more than 2500 topic-segmented sequences. Speakers with different language mastery provided data in three other languages at different durations and in pairs. The ratio of actors/actresses, range of phonemes, and variety of languages are carefully designed to cover natural language characteristics. For emotional gestures, feedback on the speakers' expressions was provided by professional instructors during the recording process and re-recorded in case of non-expressive gesturing to ensure the expressiveness and quality of the entire dataset. After statistical analysis on BEAT, we observed the correlation of conversational gestures with *facial expressions*, *emotions*, and *semantics*, in addition to the known correlation with *audio*, *text*, and *speaker identity*.

Additionally, we propose a baseline neural network architecture, **Ca**scaded **M**otion **N**etwork (**CaMN**), which learns synthesizing body and hand gestures by inputting all six modalities mentioned above. The proposed model consists of cascaded encoders and decoders for enhancing the contribution of audio and facial modalities. Besides, in order to evaluate the semantic relevancy, we propose **S**emantic-**R**elevant **G**esture **R**ecall (**SRGR**), which weights Probability of Correct Keypoint (PCK) based on semantic scores of the ground truth data. Overall, our contributions can be summarized as follows:

– We release BEAT, which is the first gesture dataset with semantic and emotional annotation, and the largest motion capture dataset in terms of duration and available modalities to the best of our knowledge.
– We propose CaMN as a baseline model that inputs audio, text, facial blendweight, speaker identity, emotion and semantic score to synthesize conversational body and hand gestures through cascaded network architecture.
– We introduce SRGR to evaluate the semantic relevancy as well as the human preference for conversational gestures.

Finally, qualitative and quantitative experiments demonstrate the data quality of BEAT, the state-of-the-art performance of CaMN and the validness of SRGR.

2   Related Work

Conversational Gestures Dataset. We first review mo-cap and pseudo-label conversational gestures datasets. Volkova *et al.* [47] built a mo-cap emotional gestures dataset in 89 mins with text annotation, Takeuchi *et al.* [45] captured an interview-like audio-gesture dataset in total 3.5-hour with two Japanese speakers. Ferstl and Mcdonnell [17] collected a 4-hour dataset, Trinity, with a single male speaker discussing hobbies, *etc.*, which is the most common used mo-cap dataset for conversational gestures synthesis. On the other hand, Ginosar *et al.* [20] used OpenPose [12] to extract 2D poses from YouTube videos as training data for 144 hours, called S2G Dataset. Habibie *et al.* [21] extended it to a full 3D body with facial landmarks, and the last available data is 33 hours. Similarly, Yoon *et al.* [52] used VideoPose3D [39] to build on the TED dataset, which is 97 hours with 9 joints on upper body. The limited data amount of mo-cap and noise in ground truth makes a trade-off for the trained network's generalization capability and quality. Similar to our work, several datasets are built for talking-face generation and the datasets can be divided into 3D scan face, *e.g.*, VOCA [46] and MeshTalk [42] or RGB images [4,11,15,26,49]. However, these datasets cannot be adopted to synthesize human gestures.

Semantic or Emotion-Aware Motion Synthesis. Semantic analysis of motion has been studied in the action recognition and the sign-language analysis/synthesis research domains. For example, in some of action recognition datasets [25,28,13,14,43,9,34,44,40,48] clips of action with the corresponding label of a single action, *e.g.*, running, walking [41] is used. Another example is audio-driven sign-language synthesis [27], where hand gestures have specific semantics. However, these datasets do not apply to conversational gestures synthesis since gestures used in natural conversations are more complex than single actions, and their semantic meaning differs from sign-language semantics. Recently, Bhattacharya [7] extracted emotional cues from text and used them for gesture synthesis. However, the proposed method has limitations in the accuracy of the emotion classification algorithm and the diversity of emotion categories in the dataset.

Conditional Conversational Gestures Synthesis. Early baseline models were released with datasets such as text-conditioned gesture [53], audio-conditioned gesture [20,45,17], and audio-text-conditioned gesture [52]. These baseline models were based on CNN and LSTM for end-to-end modelling. Several efforts try to improve the performance of the baseline model by input/output representation selection [30,19], adversarial training [18] and various types of generative modeling techniques [50,36,51,1], which can be summarized by "Estimating a better distribution of gestures based on the given conditions.". As an example, StyleGestures [2] uses Flow-based model [23] and additional control signal to sample gesture from the distribution. Probabilistic gesture generation enables generating diversity based on noise, which is achieved by CGAN [51], WGAN [50]. However, due to the lack of paired multi-modal data, the analysis of other modalities, *e.g.*, facial expression, for gesture synthesis is still missing.

3   BEAT: Body-Expression-Audio-Text Dataset

In this section, we introduce the proposed Body-Expression-Audio-Text (BEAT) Dataset. First, we describe the dataset acquisition process and then introduce text, emotion, and semantic relevance information annotation. Finally, we use BEAT to analyze the correlation between conversational gestures and emotions and show the distribution of semantic relevance.

3.1   Data Acquisition
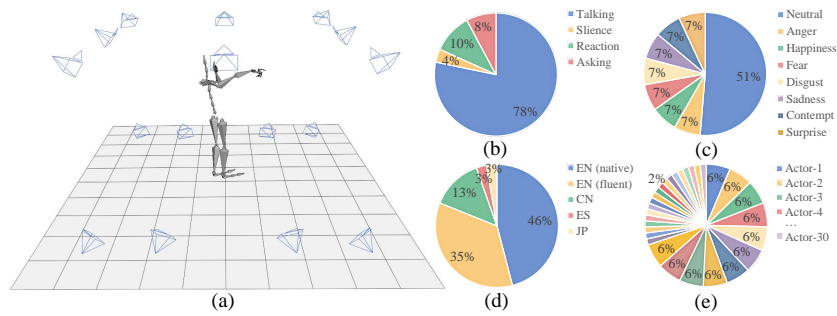
Fig. 2. Capture System and Subject Distribution of BEAT. (a) A 16-camera motion capture system is adopted to record data in Conversation and Self-Talk sessions. (b) Gestures are divided into four categories in Conversation session. (c) Seven additional emotion categories are set in equal proportions in the self-talk session. Besides, (d) our dataset includes four languages which mainly consist of English, (e) by 30 speakers from ten countries with different recording duration.

Motion Capture System. The motion capture system shown in Figure 2a, is based on 16 synchronized cameras recording motion at 120 Hz. We use Vicon's suits with 77 markers (*cf.* supplementary materials for the location of markers on the body). The facial capture system uses ARKit with a depth camera on iPhone 12 Pro, which extracts 52 blendshape weights at 60 Hz. The blendshape targets are designed based on Facial Action Coding System (FACS) and are widely used by industry novice users. The audio is recorded in a 48KHz stereo.

Design Criteria. BEAT is equally divided into *conversation* and *self-talk* sessions, which consist of 10-min and 1-min sequences, respectively. The conversation is between the speaker and the instructor remotely, *i.e.*, to ensure only the speaker's voice is recorded. As shown in Figure 2b, The speaker's gestures are divided into four categories talking, instantaneous reactions to questions, the state of thinking (silence) and asking. We timed each category's duration during the recording process. Topics were selected from 20 predefined topics, which cover 33% and 67% debate and description topics, respectively. *Conversation* sessions would record the neutral conversations without acting to ensure the diversity of the dataset. The *self-talk* sessions consist of 120 1-minute self-talk recordings,

where speakers answer questions about daily conversation topics, *e.g.*, personal experiences or hobbies. The answers were written and proofread by three English native speakers, and the phonetic coverage was controlled to be similar to the frequently used 3000 words [24]. We covered 8 emotions, *neutral, anger, happiness, fear, disgust, sadness, contempt* and *surprise*, in the dataset referring to [35] and the ratio of each emotion is shown in Figure 2c. Among the 120 questions, 64 were for neutral emotions, and the remaining seven had eight questions each. Different speakers were asked to talk about the same content with their personalized gestures. Details about predefined answers and pronunciation distribution are available in the supplementary materials.

Speaker Selection and Language Ratio. We strictly control the proportion of languages as well as accents to ensure the generalization capability of the dataset. As shown in Figure 2d, the dataset consists mainly of English data: 60h (81%), 12h of Chinese, 2h of Spanish and Japanese. The Spanish and Japanese are also 50% of the size of the previous mo-cap dataset [17]. The English component includes 34h of 10 native English speakers, including the US, UK, and Australia, and 26h of 20 fluent English speakers from other countries. As shown in Figure 2e, 30 speakers (including 15 females) from different ethnicities can be grouped into two depending on their total recording duration as 4-hour (10 speakers) and 1-hour (20 speakers), where the 1-hour data is proposed for few-shot learning experiments. It is recommended to check the supplementary material for details of the speakers.

Recording. Speakers were asked to read answers in self-talk sections proficiently. However, they were not guided to perform a specific style of gesture but were encouraged to show a natural, personal, daily style of conversational gestures. Speakers would watch 2-10 minutes of emotionally stimulating videos corresponding to different emotions before talking with the particular emotion. A professional speaker would instruct them to elicit the corresponding emotion correctly. We re-record any unqualified data to ensure the data's correctness and quality.

3.2   Data Annotation

Text Alignment. We use an in-house-built Automatic Speech Recognizer (ASR) to obtain the initial text for the conversation session and proofread it by annotators. Then, we adopt Montreal Forced Aligner (MFA) aligner [37] for temporal alignment of the text with audio.

Emotion and Semantic Relevance. The 8-class emotion label of self-talk is confirmed, and the on-site supervision guarantees the correctness. For the conversation session, annotators would watch the video with corresponding audio and gestures to perform frame-level annotation. For the semantic relevance, we get the score on a scale of 0-10 from assigned 600 annotators from Amazon Mechanical Turk (AMT). The annotators were asked to annotate a small amount of test data as a qualification check, of which only 118 annotators succeeded in the qualification phase for the final data annotation. We paid $\sim$ \$10 for each annotator per hour in this task.

Fig. 3. Emotional Gesture Clustering and Examples. (a) T-SNE visualization for gestures in eight emotion categories. Gestures with different emotions are basically distinguished into different groups, *e.g.*, the Happiness (blue) and Anger (orange). (b) Examples of Happiness (top) and Anger gestures from speaker-2.

3.3   Data Analysis

The collection and annotation of BEAT have made it possible to analyze correlations between conversational gestures and other modalities. While the connection between gestures and audio, text and speaker identity has been widely studied. We further discuss the correlations between gestures, facial expressions, emotions, and semantics.

Facial Expression and Emotion. Facial expressions and emotions were strongly correlated (excluding some of the lip movements), and we first analyze the correlation between conversational gestures and emotional categories here. As shown in Figure 3a, We visualized the gestures in T-SNE based on a 2s-rotation representation, and the results showed that gestures have different characteristics in different emotions. For example, as shown in Figure 3b, speaker-2 has different gesture styles when angry and happy, *e.g.*, the gestures are larger and faster when angry. The T-SNE results also significantly differ between happy (blue) and angry (yellow). However, the gestures for the different emotions are still not perfectly separable by the rotation representation. Furthermore, the gestures of the different emotions appear to be confounded in each region, which is also consistent with subjective perceptions.

Distribution of Semantic Relevance. There is large randomness for the semantic relevance between gestures and texts, which is shown in Figure 4, where the frequency, position and content of the semantic-related gestures vary from speaker to speaker when the same text content is uttered. In order to better understand the distribution of the semantic relevance of the gestures, we conducted a semantic relevance study based on four hours of two speakers' data. As shown in Figure 4b, for the overall data, 83% of the gestures have low semantic scores (≤ 0.2). For the words-level, the semantic distribution varied between words, *e.g.*, *i* and *was* which are sharing a similar semantic score but different in the score distribution. Besides, Figure 4c shows the average semantic scores of nine high-frequency words in the text corpus. It is to be mentioned that the scores of
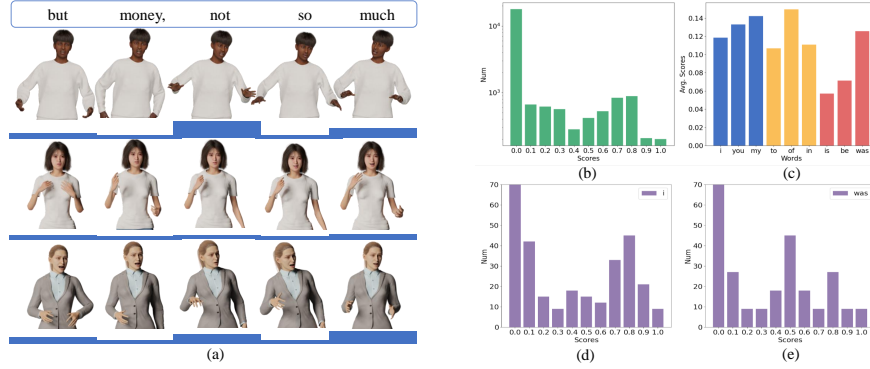
Fig. 4. Distribution of semantic labels. (a) Different speaker ID speaks in a same phase happens different levels of semantic relevance and different styles of gesture. (b) The overall semantic distribution of BEAT. (c) The semantic relevance of the high frequency words which are grouped by their lexical in different color. (d, e) Different distribution of semantic relevance happens in words *i* and *was* even sharing almost the same level of semantic relevance.

the *Be-verbs* showed are comparatively lower than that *Pronouns* and *Prepositions* which are shown in blue and yellow, respectively. Ultimately, it presents a different probability distribution to the semantically related gestures.

## 4   Multi-Modal Conditioned Gestures Synthesis Baseline

In this section, we propose a baseline that inputs all the modalities for generating vivid, human-like conversational gestures. The proposed baseline, Cascaded Motion Network (CaMN), is shown in Figure 5, which encodes text, emotion condition, speaker identity, audio and facial blendshape weights to synthesize body and hands gestures in a multi-stage, cascade structure. In addition, semantic relevancy is adopted as a loss weight to make the network generate more semantic-relevant gestures. The text, audio and speaker ID encoders network selection are referred to [52] and customized for better performance. All input data have the same time resolution as the output gestures so that the synthesized gestures can be processed frame by frame through a sequential model. The gesture and facial blendshape weights are downsampled to 15 FPS, and the word sentence is inserted with padding tokens to correspond to the silence time in the audio.

Text Encoder. First, words are converted to word embedding set $\mathbf{v}^T \in \mathbb{R}^{300}$ by pre-trained model in FastText [10] to reduce dimensions. Then, the word sets are fine-tuned by customized encoder E_T, which is a 8-layer temporal convolution network (TCN) [6] with skip connections [22], as

$$z_i^T = E_T(v_{i-f}^T, ..., v_{i+f}^T), \tag{1}$$

For each frame i, the TCN fusions the information from $2f = 34$ frames to generate final latent feature of text, the set of features is note as $\mathbf{z}^T \in \mathbb{R}^{128}$.
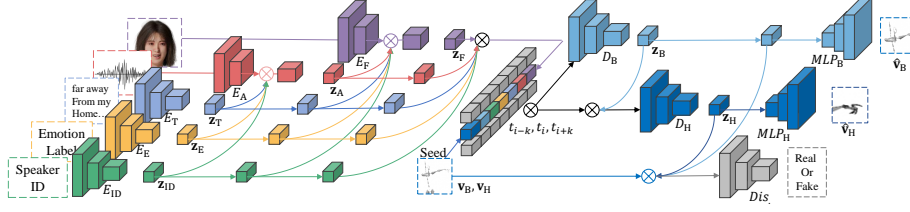
Fig. 5. Cascaded Motion Network (CaMN). As a multi-modal gesture synthesis baseline, CaMN inputs *text, emotion label, speaker ID, audio* and *facial blendweight* in a cascaded architecture, the audio and facial feature will be extracted by concatenating the features of previous modalities. The fused feature will be reconstructed to body and hands gestures by two cascaded LSTM+MLP decoders.

Speaker ID and Emotion Encoders. The initial representation of speaker ID and emotion are both one-hot vectors, as $\mathbf{v}^{\mathrm{ID}} \in \mathbb{R}^{30}$ and $\mathbf{v}^{\mathrm{E}} \in \mathbb{R}^{8}$. Follow the suggestion in [52], we use embedding-layer as speaker ID encoder, E_{ID}. As the speaker ID does not change instantly, we only use the current frame speaker ID to calculate its latent features. On the other hand, we use a combination of embedding-layer and 4-layer TCN as the emotion encoder, E_{E}, to extract the temporal emotion variations.

$$z_i^{\mathrm{ID}} = E_{\mathrm{ID}}(v_i^{\mathrm{ID}}), z_i^{\mathrm{E}} = E_{\mathrm{E}}(v_{i-f}^{\mathrm{E}}, ..., v_{i+f}^{\mathrm{E}}), \tag{2}$$

where $\mathbf{z}^{\mathrm{ID}} \in \mathbb{R}^{8}$ and $\mathbf{z}^{\mathrm{E}} \in \mathbb{R}^{8}$ is the latent feature for speaker ID and emotion, respectively.

Audio Encoder. We adopt the raw wave representation of audio and downsample it to 16KHZ, considering audio as 15FPS, for each frame, we have $\mathbf{v}^{\mathrm{A}} \in \mathbb{R}^{1067}$. We feed the audio joint with the text, speakerID and emotion features into audio encoder E_{A} to learn better audio features. As

$$z_i^{\mathrm{A}} = E_{\mathrm{A}}(v_{i-f}^{\mathrm{A}}, ..., v_{i+f}^{\mathrm{E}}; v_i^{\mathrm{T}}; v_i^{\mathrm{E}}; v_i^{\mathrm{ID}}), \tag{3}$$

The E_{A} consists of 12-layer TCN with skip connection and 2-layer MLP, features in other modifies are concatenated with the 12th layer audio features thus the final MLP layers are for audio feature refinement, and the final latent audio feature is $\mathbf{z}^{\mathrm{A}} \in \mathbb{R}^{128}$.

Facial Expression Encoder. We take the $\mathbf{v}^{\mathrm{F}} \in \mathbb{R}^{52}$ as initial representation of facial expression. 8-layer TCN and 2-layer MLP based encoder E_{F} is adopt to extract facial latent feature $\mathbf{z}^{\mathrm{F}} \in \mathbb{R}^{32}$, as

$$z_i^{\mathrm{F}} = E_{\mathrm{F}}(v_{i-f}^{\mathrm{F}}, ..., v_{i+f}^{\mathrm{F}}; v_i^{\mathrm{T}}; v_i^{\mathrm{E}}; v_i^{\mathrm{ID}}; v_i^{\mathrm{A}}), \tag{4}$$

the features are concatenated at 8th layer and the MLP is for refinement.

Body and Hands Decoders. We implement the body and hands decoders in a separated, cascaded structure, which is based on [38] conclusion that the body gestures can be used to estimate hand gestures. These two decoders, D_{B}

and D_{F} are based on the LSTM structure for latent feature extraction and 2-layer MLP for gesture reconstruction. They would combine the features of five modalities with previous gestures, $i.e.$, seed pose, to synthesis latent gesture features $\mathbf{z}^{\mathrm{B}} \in \mathbb{R}^{256}$ and $\mathbf{z}^{\mathrm{H}} \in \mathbb{R}^{256}$. The final estimated body $\hat{\mathbf{v}}^{\mathrm{B}} \in \mathbb{R}^{27 \times 3}$ and hands $\hat{\mathbf{v}}^{\mathrm{H}} \in \mathbb{R}^{48 \times 3}$ are calculated as,

$$z_i^{\mathrm{M}} = z_i^{\mathrm{T}} \otimes z_i^{\mathrm{ID}} \otimes z_i^{\mathrm{E}} \otimes z_i^{\mathrm{A}} \otimes z_i^{\mathrm{F}} \otimes v_i^{\mathrm{B}} \otimes v_i^{\mathrm{H}}, \tag{5}$$

$$\mathbf{z}^{\mathrm{B}} = D_{\mathrm{B}}(z_0^{\mathrm{M}}, ..., z_n^{\mathrm{M}}), \mathbf{z}^{\mathrm{H}} = D_{\mathrm{H}}(z_0^{\mathrm{M}}, ..., z_n^{\mathrm{M}}; \mathbf{z}^{\mathrm{B}}), \tag{6}$$

$$\hat{\mathbf{v}}^{\mathrm{B}} = MLP_{\mathrm{B}}(\mathbf{z}^{\mathrm{B}}), \hat{\mathbf{v}}^{\mathrm{H}} = MLP_{\mathrm{H}}(\mathbf{z}^{\mathrm{H}}), \tag{7}$$

$\mathbf{z}^{\mathrm{M}} \in \mathbb{R}^{549}$ is the merged features for all modalities. For Equation 5, the length for the seed pose is four frames.

Loss Functions. The final supervision of our network is based on gesture reconstruction and the adversarial loss

$$\ell_{\mathrm{Gesture\ Rec.}} = \mathbb{E}\left[\left\|\mathbf{v}^B - \hat{\mathbf{v}}^B\right\|_1\right] + \alpha \mathbb{E}\left[\left\|\mathbf{v}^H - \hat{\mathbf{v}}^H\right\|_1\right], \tag{8}$$

$$\ell_{\mathrm{Adv.}} = -\mathbb{E}[\log(Dis(\hat{\mathbf{v}}^B; \hat{\mathbf{v}}^H))], \tag{9}$$

where the discriminator input to the adversarial training is only the gesture itself. We also adopt a weight α to balance the body and hands penalties. After that, during training, we adjust the weights of L1 loss, and adversarial loss using the semantic-relevancy label λ The final loss function is

$$\ell = \lambda \beta_0 \ell_{\mathrm{Gesture\ Rec.}} + \beta_1 \ell_{\mathrm{Adv}}, \tag{10}$$

where β_0 and β_1 are predefined weight for L1 and adversarial loss. When semantic relevancy is high, we encourage the network to generate gestures spatially similar to ground truth as much as possible, thus strengthening the L1 penalty and decreasing the adversarial penalty.

5 Metric for Semantic Relevancy

We propose the Semantic-Relevant Gesture Recall (SRGR) to evaluate the semantic relevancy of gestures, which can also be interpreted as whether the gestures are vivid and diverse. We utilize the semantic scores as a weight for the Probability of Correct Keypoint (PCK) between the generated gestures and the ground truth gestures. Where PCK is the number of joints successfully recalled against a specified threshold δ. The SRGR metric can be calculated as follows:

$$D_{SRGR} = \lambda \sum \frac{1}{T \times J} \sum_{t=1}^{T} \sum_{j=1}^{J} \mathbf{1}\left[\left\|p_t^j - \hat{p}_t^j\right\|_2 < \delta\right], \tag{11}$$

where $\mathbf{1}$ is the indicator function and T, J is the set of frames and number of joints. We think the SRGR, which emphasizes recalling gestures in the clip of interest, is more in line with the subjective human perception of gesture's valid diversity than the L1 variance of synthesized gestures.

6   Experiments

In this section, we first evaluate the SRGR metric's validity, then demonstrate our dataset's data quality based on subjective experiments. Next, we demonstrate the validity of our baseline model using subjective and objective experiments, and finally, we discuss the contribution of each modality based on ablation experiments.

6.1   Validness of SRGR
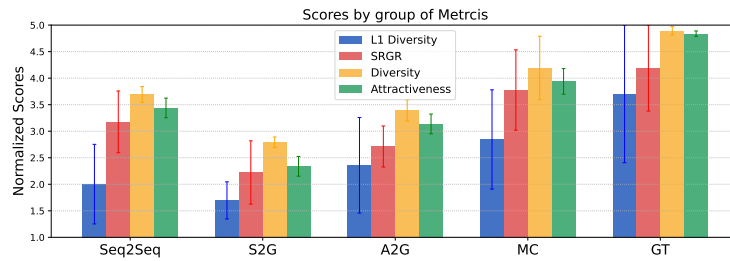
Fig. 6. Comparison of Metrics by Group. SRGR shows the consistence with human perception, and lower variance comparing with L1 Diversity in evaluation.

A user study is conducted to evaluate the validity of SRGR. Firstly, we randomly trim the motion sequences with rendered results into clips which are around 40s. For each clip, the participants are asked to evaluate the gesture based on its diversity which is the number of non-repeated gestures. Besides, the participants then need to score its attractiveness which should be based on the motion itself instead of the content of the speech. Totally 160 participants took part in the evaluation study, and each of them evaluated 15 random clips of gestures. There are totally 200 gesture clips including the results generated by using the methods from Seq2Seq [53], S2G [20], A2G [32], MultiContext [52], and ground truth, 40 clips for each with the same speaker data. Both of the questions follow a 5-points Likert scale. As shown in Figure 6, we found a large variance in L1 diversity even though we used 100 gesture segments to calculate the average L1 distance, (usually around 40 segments [33,32]). Secondly, generated results with strong semantic relevance but a smaller motion range, such as Seq2Seq, obtained a lower L1 diversity than A2G, which has a larger motion range, yet the statistical evidence that humans feel that Seq2Seq has higher diversity than A2G. An explanation is a human evaluating diversity not only on the range of motion but also on some other implicit features, such as expressiveness and semantic relevancy of the motion.

6.2   Data quality

To evaluate the captured ground truth motion data quality, we compare our proposed dataset with the widely used mocap dataset Trinity [17] and in-the-wild

dataset S2G-3D [20,21]. We conducted the user study by comparing clips sampled from ground truth and generated results using motion synthesis networks trained in each dataset. The Trinity dataset has a total of 23 sequences, with 10 minutes each. We randomly divide the data into 19:2:2 for train/valid/test since there is no standard for splitting.

Table 2. User Study Comparison with Trinity for Data Quality. Comparing with Trinity [17], BEAT get higher user preference score in terms of ground truth data quality. "-b" and "-h" indicate body and hands, respectively.

	correctness-b			correctness-h			diversity			synchrony		
	S2G	A2G	GT	S2G	A2G	GT	S2G	A2G	GT	S2G	A2G	GT
Trinity [17]	38.8	37.0	43.8	15.3	14.6	11.7	42.1	36.7	40.2	40.9	36.3	46.4
BEAT (Ours)	61.2	63.0	56.2	84.7	85.4	88.3	57.9	63.3	59.8	59.1	63.7	53.6

Table 3. User Study Comparison with S2G-3D. BEAT get similar user preferences in terms of naturalness. Based on the score, the model trained on BEAT dataset would be fitted into a more physically correct, diverse, and attractive distribution.

	naturalness	correctness	diversity	attractiveness
S2G-3D [20,21]	33.03 ± 1.93	21.17 ± 2.84	29.17 ± 1.81	28.79 ± 2.53
BEAT (conversation)	34.16 ± 2.16	39.94 ± 3.97	34.69 ± 1.76	29.90 ± 2.19
BEAT (self-talk)	32.81 ±1.79	38.89 ± 3.75	36.14 ± 1.99	42.31 ± 2.40

We used S2G [20], as well as the SoTA algorithm A2G [32], to cover both GAN and VAE models. The output layer of the S2G model was adapted for outputting 3D coordinates. In the ablation study, the final generated 3D skeleton results were rendered and composited with audio for comparison in the user study. A total of 120 participant subjects compared the clips randomly sampled from Trinity and our dataset, with 5-20s in length. The participants were asked to evaluate gestures correctness, *i.e.*, physical correctness, diversity and gesture-audio synchrony. Furthermore, the body and hands were evaluated separately for the gesture correctness test. The results are shown in Table 2, demonstrating that our dataset received higher user preference in all aspects. Especially for the hand movements, we outperformed the Trinity dataset by a large margin. This is probably due to the noise of the past motion capture devices and the lack of markers on the hands. Table 3 shows preference ratios (%) of 60 subjects who watch 20 random rendered 3D skeletons pairs per subjective test. Based on the score, the model trained on the BEAT dataset would be fitted into a more physically correct, diverse, and attractive distribution.

6.3   Evaluation of the baseline model

Training Setting. We use the Adam optimizer [29] to train at a learning rate of 2e-4, and the 4-speaker data is trained in an NVIDIA V100 environment. For evaluation metrics, L1 has been demonstrated unsuitable for evaluating the gesture performance [32,52] thus, we adopt FGD [52] to evaluate the generated

Table 4. Evaluation on BEAT. Our CaMN performs best in the term of FGD, SRGR and BeatAlign, all methods are trained on our dataset (BEAT)

	FGD ↓	SRGR ↑	BeatAlign ↑
Seq2Seq [53]	261.3	0.173	0.729
S2G [20]	256.7	0.092	0.751
A2G [32]	223.8	0.097	0.766
MultiContext [52]	176.2	0.196	0.776
CaMN (Ours)	123.7	0.239	0.783

Table 5. Results of Ablation Study.

	FGD ↓	BGSR ↑	BeatAlign ↑
full cascaded	123.7	0.239	0.783
w/o cascaded	137.2	0.207	0.776
w/o text	149.4	0.171	0.781
w/o audio	155.7	0.225	0.733
w/o speaker ID	159.1	0.207	0.774
w/o face	163.3	0.217	0.767
w/o emotion	151.7	0.231	0.775
w/o semantic	151.8	0.194	0.786

gestures' distribution distance with ground truth. It computes the distance between latent features extracted by a pretrained network, we use an LSTM-based autoencoder as the pretrained network. In addition, we adopt SRGR and Beat-Align to evaluate diversity and synchrony. BeatAlign [33] is a Chamfer Distance between audio and gesture beats to evaluate gesture-audio beat similarity.

Quantitative Results. The final results are shown in Table 4. In addition to S2G and A2G, we also compare our results with text-to-gesture and audio&test-to-gesture algorithm, Seq2Seq [53] and MultiContext [52]. The results show that both our end2end model and cascaded model archive SoTA performance in all metrics (*cf.* supplementary materials for video results).

6.4   Ablation Study.

Effectiveness of Cascaded Connection. As shown in Table 5, in contrast to the end-to-end approach, the cascaded connection can achieve better performance because we introduce prior human knowledge to help the network extract features of different modalities.

Effectiveness of Each Modality. We gradually removed the data of one modality during the experiment (*cf.* Table 5). Synchrony would significantly be reduced after removing the audio, which is intuitive. However, it still maintains some synchronizations, such as the padding and time-align annotation of the text and the lip motion of the facial expression. In contrast, eliminating weighted semantic loss improves synchrony, which means that semantic gestures are usually not strongly aligned with audio perfectly. There is also a relationship between emotion and synchrony, but speaker ID only has little effect on synchrony. The removal of audio, emotion, and facial expression does not significantly affect the semantic relevant gesture recall, which depends mainly on the text and the speaker ID. Data from each modality contributed to improving the FGD, which means using different modalities of data enhances the network's mapping ability. The unities of audio and facial expressions, especially facial expressions, improve the FGD significantly. We found that removing emotion and speaker ID also impacts the FGD scores. This is because using the integrated network increases the diversity of features, which leads to a diversity of results, increasing the variance of the distribution and making it more like the original data.

Emotional Gestures. As shown in Table 6, we train a classifier by an additional 1DCNN + LSTM network and invite 60 subjects each to classify 12 random real test clips (with audio). The classifier is trained and tested on speaker-4's ground truth data.

Table 6. Emotional Gesture Classification. The classification accuracy (%) gap between the test real and generated data (1344 clips, 10s each) is 15.85.

	neutral	happiness	sadness	anger	surprise	contempt	fear	disgust	avg.
human	84.29	74.86	82.65	88.36	76.12	71.59	80.94	72.33	78.89
real	51.06	98.68	85.08	38.78	99.39	81.08	99.95	99.62	83.26
generated	36.95	76.83	62.17	37.46	77.91	70.61	81.32	83.03	67.41

6.5   Limitation

Impact of Acting. Self-Talk sessions might reflect the impact of acting, which is inevitable and controlled. *Inevitable:* The impact is probably caused by pre-defined content. However, to explore the semantic-relevancy and personality, it is necessary to control the variables, *i.e.*, different speakers should talk in the same text and emotion so that the personality can be carefully explored. *Controlled.* Speakers recorded the conversation session first and were encouraged to keep the same style as the conversation. We also filtered out about 21h of data and six speakers due to inconsistencies in their styles.

Calculation of SRGR. SRGR now is calculated based on semantic annotation, which has a limitation for an un-labelled dataset. To solve this problem, training a scoring network or semantic discriminator is a possible direction.

7   Conclusion

We build a large-scale, high-quality, multi-modal, semantic and emotional annotated dataset to generate more human-like, semantic and emotional relevant conversational gestures. Together with the dataset, we propose a cascade-based baseline model for gesture synthesis based on six modalities and achieve SoTA performance. Finally, we introduce SRGR for evaluating semantic relevancy. In the future, we plan to expand cross-data checks for AU and emotion recognition benchmarks. Our dataset and the related statistical experiments could benefit a number of different research fields, including controllable gesture synthesis, cross-modality analysis and emotional motion recognition in the future.

8   Acknowledgements

References

1. Ahuja, C., Lee, D.W., Nakano, Y.I., Morency, L.P.: Style transfer for co-speech gesture animation: A multi-speaker conditional-mixture approach. In: European Conference on Computer Vision. pp. 248–265. Springer (2020) 4

2. Alexanderson, S., Henter, G.E., Kucherenko, T., Beskow, J.: Style-controllable speech-driven gesture synthesis using normalising flows. In: Computer Graphics Forum. vol. 39, pp. 487–496. Wiley Online Library (2020) 4

3. Alexanderson, S., Székely, É., Henter, G.E., Kucherenko, T., Beskow, J.: Generating coherent spontaneous speech and gesture from text. In: Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents. pp. 1–3 (2020) 2

4. Alghamdi, N., Maddock, S., Marxer, R., Barker, J., Brown, G.J.: A corpus of audiovisual lombard speech with frontal and profile views. The Journal of the Acoustical Society of America **143**(6), EL523–EL529 (2018) 4

5. Ali, G., Lee, M., Hwang, J.I.: Automatic text-to-gesture rule generation for embodied conversational agents. Computer Animation and Virtual Worlds **31**(4-5), e1944 (2020) 2

6. Bai, S., Kolter, J.Z., Koltun, V.: An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. arXiv preprint arXiv:1803.01271 (2018) 8

7. Bhattacharya, U., Childs, E., Rewkowski, N., Manocha, D.: Speech2affectivegestures: Synthesizing co-speech gestures with generative adversarial affective expression learning. In: Proceedings of the 29th ACM International Conference on Multimedia. pp. 2027–2036 (2021) 2, 4

8. Bhattacharya, U., Rewkowski, N., Banerjee, A., Guhan, P., Bera, A., Manocha, D.: 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. In: 2021 IEEE Virtual Reality and 3D User Interfaces (VR). pp. 1–10. IEEE (2021) 2

9. Bloom, V., Makris, D., Argyriou, V.: G3d: A gaming action dataset and real time action recognition evaluation framework. In: 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops. pp. 7–12. IEEE (2012) 4

10. Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching word vectors with subword information. Transactions of the association for computational linguistics **5**, 135–146 (2017) 8

11. Cao, H., Cooper, D.G., Keutmann, M.K., Gur, R.C., Nenkova, A., Verma, R.: Crema-d: Crowd-sourced emotional multimodal actors dataset. IEEE transactions on affective computing **5**(4), 377–390 (2014) 4

12. Cao, Z., Hidalgo, G., Simon, T., Wei, S.E., Sheikh, Y.: Openpose: realtime multi-person 2d pose estimation using part affinity fields. IEEE transactions on pattern analysis and machine intelligence **43**(1), 172–186 (2019) 2, 4

13. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6299–6308 (2017) 4

14. Chen, C., Jafari, R., Kehtarnavaz, N.: Utd-mhad: A multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor. In: 2015 IEEE International conference on image processing (ICIP). pp. 168–172. IEEE (2015) 4

15. Cooke, M., Barker, J., Cunningham, S., Shao, X.: An audio-visual corpus for speech perception and automatic speech recognition. The Journal of the Acoustical Society of America **120**(5), 2421–2424 (2006) 4

16. Cudeiro, D., Bolkart, T., Laidlaw, C., Ranjan, A., Black, M.J.: Capture, learning, and synthesis of 3d speaking styles. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10101–10111 (2019) 3

17. Ferstl, Y., McDonnell, R.: Investigating the use of recurrent motion modelling for speech gesture generation. In: Proceedings of the 18th International Conference on Intelligent Virtual Agents. pp. 93–98 (2018) 2, 3, 4, 6, 11, 12

18. Ferstl, Y., Neff, M., McDonnell, R.: Adversarial gesture generation with realistic gesture phasing. Computers & Graphics **89**, 117–130 (2020) 4

19. Ferstl, Y., Neff, M., McDonnell, R.: Expressgesture: Expressive gesture generation from speech through database matching. Computer Animation and Virtual Worlds p. e2016 (2021) 4

20. Ginosar, S., Bar, A., Kohavi, G., Chan, C., Owens, A., Malik, J.: Learning individual styles of conversational gesture. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3497–3506 (2019) 2, 3, 4, 11, 12, 13

21. Habibie, I., Xu, W., Mehta, D., Liu, L., Seidel, H.P., Pons-Moll, G., Elgharib, M., Theobalt, C.: Learning speech-driven 3d conversational gestures from video. arXiv preprint arXiv:2102.06837 (2021) 2, 3, 4, 12

22. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016) 8

23. Henter, G.E., Alexanderson, S., Beskow, J.: Moglow: Probabilistic and controllable motion synthesis using normalising flows. ACM Transactions on Graphics (TOG) **39**(6), 1–14 (2020) 4

24. Hornby, A.S., et al.: Oxford advanced learner's dictionary of current English. Oxford University Press, Oxford.[OALDCE] 6

25. Ionescu, C., Papava, D., Olaru, V., Sminchisescu, C.: 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), 1325–1339 (2013) 4

26. Jackson, P., Haq, S.: Surrey audio-visual expressed emotion (savee) database. University of Surrey: Guildford, UK (2014) 4

27. Kapoor, P., Mukhopadhyay, R., Hegde, S.B., Namboodiri, V., Jawahar, C.: Towards automatic speech to sign language generation. arXiv preprint arXiv:2106.12790 (2021) 4

28. Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., et al.: The kinetics human action video dataset. arXiv preprint arXiv:1705.06950 (2017) 4

29. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014) 12

30. Kucherenko, T., Hasegawa, D., Kaneko, N., Henter, G.E., Kjellström, H.: Moving fast and slow: Analysis of representations and post-processing in speech-driven automatic gesture generation. International Journal of Human–Computer Interaction pp. 1–17 (2021) 4

31. Kucherenko, T., Jonell, P., van Waveren, S., Henter, G.E., Alexandersson, S., Leite, I., Kjellström, H.: Gesticulator: A framework for semantically-aware speech-driven gesture generation. In: Proceedings of the 2020 International Conference on Multimodal Interaction. pp. 242–250 (2020) 2

32. Li, J., Kang, D., Pei, W., Zhe, X., Zhang, Y., He, Z., Bao, L.: Audio2gestures: Generating diverse gestures from speech audio with conditional variational autoencoders. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 11293–11302 (2021) 2, 11, 12, 13

33. Li, R., Yang, S., Ross, D.A., Kanazawa, A.: Ai choreographer: Music conditioned 3d dance generation with aist++. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13401–13412 (2021) 11, 13

34. Liu, A.A., Xu, N., Nie, W.Z., Su, Y.T., Wong, Y., Kankanhalli, M.: Benchmarking a multimodal and multiview and interactive dataset for human action recognition. IEEE Transactions on cybernetics 47(7), 1781–1794 (2016) 4

35. Livingstone, S.R., Russo, F.A.: The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english. PloS one 13(5), e0196391 (2018) 6

36. Lu, J., Liu, T., Xu, S., Shimodaira, H.: Double-dcccae: Estimation of body gestures from speech waveform. In: ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 900–904. IEEE (2021) 4

37. McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., Sonderegger, M.: Montreal forced aligner: Trainable text-speech alignment using kaldi. In: Interspeech. vol. 2017, pp. 498–502 (2017) 6

38. Ng, E., Ginosar, S., Darrell, T., Joo, H.: Body2hands: Learning to infer 3d hands from conversational gesture body dynamics. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11865–11874 (2021) 9

39. Pavllo, D., Feichtenhofer, C., Grangier, D., Auli, M.: 3d human pose estimation in video with temporal convolutions and semi-supervised training. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7753–7762 (2019) 4

40. Perera, A.G., Law, Y.W., Ogunwa, T.T., Chahl, J.: A multiviewpoint outdoor dataset for human action recognition. IEEE Transactions on Human-Machine Systems 50(5), 405–413 (2020) 4

41. Punnakkal, A.R., Chandrasekaran, A., Athanasiou, N., Quiros-Ramirez, A., Black, M.J.: Babel: Bodies, action and behavior with english labels. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 722–731 (2021) 4

42. Richard, A., Zollhöfer, M., Wen, Y., De la Torre, F., Sheikh, Y.: Meshtalk: 3d face animation from speech using cross-modality disentanglement. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1173–1182 (2021) 4

43. Singh, S., Velastin, S.A., Ragheb, H.: Muhavi: A multicamera human action video dataset for the evaluation of action recognition methods. In: 2010 7th IEEE International Conference on Advanced Video and Signal Based Surveillance. pp. 48–55. IEEE (2010) 4

44. Song, S., Lan, C., Xing, J., Zeng, W., Liu, J.: An end-to-end spatio-temporal attention model for human action recognition from skeleton data. In: Proceedings of the AAAI conference on artificial intelligence. vol. 31 (2017) 4

45. Takeuchi, K., Hasegawa, D., Shirakawa, S., Kaneko, N., Sakuta, H., Sumi, K.: Speech-to-gesture generation: A challenge in deep learning approach with bidirectional lstm. In: Proceedings of the 5th International Conference on Human Agent Interaction. pp. 365–369 (2017) 4

46. Takeuchi, K., Kubota, S., Suzuki, K., Hasegawa, D., Sakuta, H.: Creating a gesture-speech dataset for speech-based automatic gesture generation. In: International Conference on Human-Computer Interaction. pp. 198–202. Springer (2017) 2, 3, 4
47. Volkova, E., De La Rosa, S., Bülthoff, H.H., Mohler, B.: The mpi emotional body expressions database for narrative scenarios. PloS one **9**(12), e113647 (2014) 3, 4
48. Wang, J., Liu, Z., Chorowski, J., Chen, Z., Wu, Y.: Robust 3d action recognition with random occupancy patterns. In: European conference on computer vision. pp. 872–885. Springer (2012) 4
49. Wang, K., Wu, Q., Song, L., Yang, Z., Wu, W., Qian, C., He, R., Qiao, Y., Loy, C.C.: Mead: A large-scale audio-visual dataset for emotional talking-face generation. In: European Conference on Computer Vision. pp. 700–717. Springer (2020) 4
50. Wu, B., Ishi, C., Ishiguro, H., et al.: Probabilistic human-like gesture synthesis from speech using gru-based wgan. In: GENEA: Generation and Evaluation of Non-verbal Behaviour for Embodied Agents Workshop 2021 (2021) 4
51. Wu, B., Liu, C., Ishi, C.T., Ishiguro, H.: Modeling the conditional distribution of co-speech upper body gesture jointly using conditional-gan and unrolled-gan. Electronics **10**(3), 228 (2021) 4
52. Yoon, Y., Cha, B., Lee, J.H., Jang, M., Lee, J., Kim, J., Lee, G.: Speech gesture generation from the trimodal context of text, audio, and speaker identity. ACM Transactions on Graphics (TOG) **39**(6), 1–16 (2020) 2, 3, 4, 8, 9, 11, 12, 13
53. Yoon, Y., Ko, W.R., Jang, M., Lee, J., Kim, J., Lee, G.: 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). pp. 4303–4309. IEEE (2019) 2, 4, 11, 13