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

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.

A Annotation Interface and Measurement of Agreement.

Our annotation interface, as shown in Figure 1a, is adapted from a modified version of VGG Image Annotator (VIA) [1] by [5]. We use the same interface for annotating both emotions and semantics. However, annotations are performed by a different group of annotators for each task. For emotion annotations, two annotators annotate the start and end times of each video segment of *Conversation* sessions. For semantics annotations, annotators would i) agree or disagree whether the current gesture is semantically related to the text content according to their perception; ii) if agreed, they annotate the start and end times for the current gesture; iii) and then, the select keyword(s) that they think the gesture exactly corresponds to from the list of keywords separated by a comma.

Our post-processing algorithm will input the separated keywords, text alignment and gesture segment-level semantic annotation to generate frame-level annotation. As shown in Figure 1b, the final semantic relevance for each frame is calculated as the multiplication of the gesture segment semantic score and the keyword semantic score. It is to be mentioned that two levels of semantic annotation can also be adopted separately.

We calculate the inter-rater agreement rate of the annotations by Measurement of inter-rater reliability for both emotional and semantic annotations. For the emotion annotations, we present scores in Table 1, the agreement is marked only two annotators give the same label. The final agreement for around 16M frames is 96%, which is high enough that we did not conduct the annotation for emotion with more than two annotators.

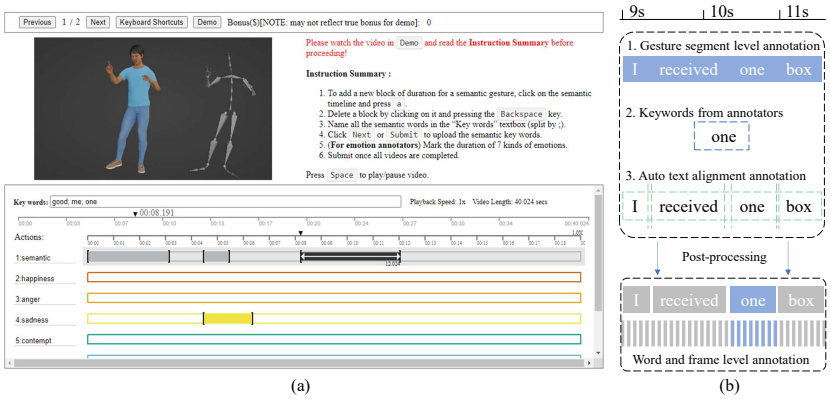
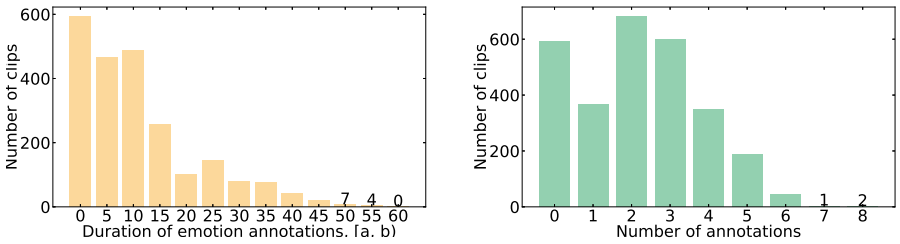


Fig. 1. Annotation Interface and Post-Processing. (a) Annotators will give the gesture segment level and keyword level annotations. (b) Post-processing algorithm will generate a frame-level semantic relevance score using segment, keyword, and text alignment annotation.

Fig. 2. Statistic on Emotion Annotation. *Left*: The sum of duration in emotion annotations (in seconds). *Right*: Distribution of annotations number in each clip.



B Details of Text Content and Speaker Information.

The distribution of vowels and consonants for the BEAT dataset is shown in Figure 3, which is basically consistent with the frequently used 3000 words [2]. For the *Conversation* sessions, the questions are selected from Table 2. There are ten topics for debate and introduction, respectively, which are related to daily conversation topics. For *Self-Talk* session, the full answer list, including 120 answers, is attached at the end of this supplementary material, which includes the translation of four languages for 30 answers proofread by native speakers.

Holding the motivation for investigating the style differences among speakers, we collected data from speakers of various countries, gender, ages and ethnicity. We then modelled these differences with explicit controls on styles. During data collection, we made sure the actor style was consistent. We filtered out about 21 hours of data and six speakers due to inconsistencies in their styles. The actor presented significantly different gestures in *Self-Talk* and *Conversation* sessions.

Table 1. Example of inter-rater reliability calculation for emotion annotation. a1 and a2 indicates annotator 1 and 2, respectively.

frame	a1	a2	agree?
01	0	0	1
02	1	2	0
03	1	1	1
04	1	0	0
avg.			0.5

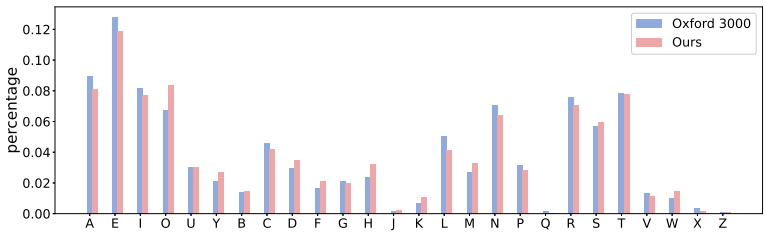


Fig. 3. Distribution of Vowel and Consonant. The distribution of our corpus is basically consistent with that of frequently used 3000 words in [2].

For example, some speakers gestured a lot during the conversation sessions but demonstrated almost no gestures in the self-talk sessions. The number of effective speakers is 30, and the corresponding recorded data duration is 76 hours. Speaker information and duration of their recordings are available in Table 3. We have 34 and 26 hours of recordings from native and fluent English speakers, respectively, and the native/fluent duration ratio is 1.307.

C Details of Data Release Format.

Our final released data file format is listed below:

i) Motion capture data in BVH file format for body and hand gestures.
ii) Audio recordings in stereo WAV file format.
iii) Facial expression blendshape weights in JSON file format.
iv) Facial mesh data in FBX file format for 8 speakers.
v) Text-audio alignment annotation data in TextGrid file format.
vi) Semantics and emotion annotations in text file format.

For body and hand gestures, the position of the markers are shown in Figure 3 (image from[1] and joint names are shown in Figure 5, respectively. We use the rotation information in Euler angles as the motion representation, *i.e.*, 75×3 rotations + 1×3 root translation.

As shown in Figure 7, facial expressions are represented with the Facial Action Coding system (FACs) based blendshapes, where each expression only acti-

[1] https://sketchfab.com/3d-models)

Table 2. List of Conversation Topics.

	Topics
1	Do you think smart-phone has destroyed communication among friends and family?
2	Now people usually work from home, for you, which one you prefer? Face-to-face communication or work from home?
3	In general, people are living longer now. Why? Discuss the causes of this phenomenon.
4	Some people believe that the Earth is being harmed (damaged) by human activity Others feel that human activity makes the Earth a better place to live. What is your opinion?.
5	Some people are always in a hurry to go places and get things done then take a rest. Other people prefer to take their time and live life at a slower pace. Which do you prefer?
6	Some people say that advertising encourages us to buy things we really do not need. Others say that advertisements tell us about new products that may improve our lives.
7	Some jobs (such as Salesmen) can be done by human or by robots (AI). Which do you prefer? get the service from human or robots (AI)?
8	Television, news, and other media pay too much attention to the personal lives of famous people such as public figures and celebrities
9	Is it more important to be able to work with a group of people on a team or to work independently?
10	Do you agree or disagree with the following statement? Only people who earn a lot of money are successful.
11	Introduce some places, cities, countries, or even planets
12	Introduce some celebrities, artists or you friends
13	Introduce some sports or physic knowledge.
14	Introduce some pets or plants.
15	Introduce some electronic Products, cars or other vehicles
16	Introduce some video games, musics, books, TV programs, stories, or movies.
17	Introduce some foods or chemical phenomenon.
18	Introduce some historical events.
19	Introduce some psychological phenomenon.
20	Introduce some military doctrine.

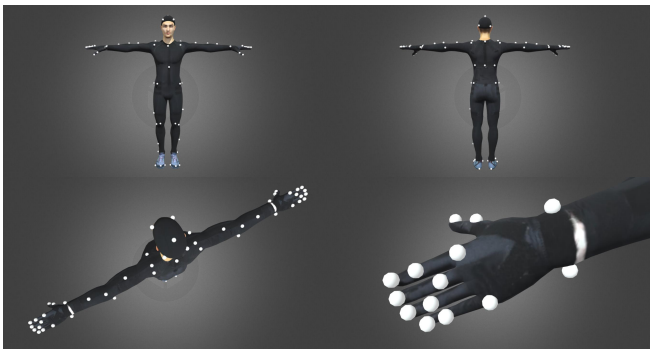


Fig. 4. Positions of motion capture markers.

Table 3. Speaker information in terms of gender, originating country, native English speaker, recording duration in English and in the second language, age and ethnicity. In the gender column, **M** and **F** stand for male and female speakers, respectively. The native column shows whether the speaker is a native English speaker. Duration is represented in hours (h).

	gender	country	native	dura.	other lan.	other dura.	age	ethnicity
1	M	US	✓	4h	-	-	25	Caucasian
2	M	US	✓	4h	-	-	32	Caucasian
3	M	US	✓	4h	-	-	40	African
4	M	Australia	✓	4h	-	-	26	Asian
5	M	UK	✓	1h	-	-	30	Caucasian
6	F	US	✓	4h	-	-	27	Caucasian
7	F	US	✓	4h	-	-	30	Caucasian
8	F	US	✓	4h	-	-	31	Asian
9	F	UK	✓	4h	-	-	32	Caucasian
10	F	UK	✓	1h	-	-	35	Caucasian
11	M	Arab	-	4h	-	-	38	African
12	M	Thailand	-	1h	-	-	32	Asian
13	M	China	-	1h	Chinese	4h	25	Asian
14	M	China	-	1h	Chinese	1h	24	Asian
15	M	China	-	1h	Chinese	1h	40	Asian
16	M	China	-	1h	-	-	32	Asian
17	M	Japan	-	1h	Japanese	1h	32	Asian
18	M	Japan	-	1h	-	-	22	Asian
19	M	Peru	-	1h	Spanish	1h	27	Caucasian
20	M	Spain	-	1h	Spanish	1h	30	Caucasian
21	F	China	-	1h	Chinese	4h	31	Asian
22	F	China	-	1h	Chinese	1h	24	Asian
23	F	China	-	1h	Chinese	1h	26	Asian
24	F	China	-	1h	-	-	32	Asian
25	F	Japan	-	1h	Japanese	1h	24	Asian
26	F	Japan	-	1h	-	-	26	Asian
27	F	Iran	-	1h	-	-	31	African
28	F	Jamaica	-	4h	-	-	33	African
29	F	Jamaica	-	1h	-	-	24	African
30	F	Russia	-	1h	-	-	25	Caucasian

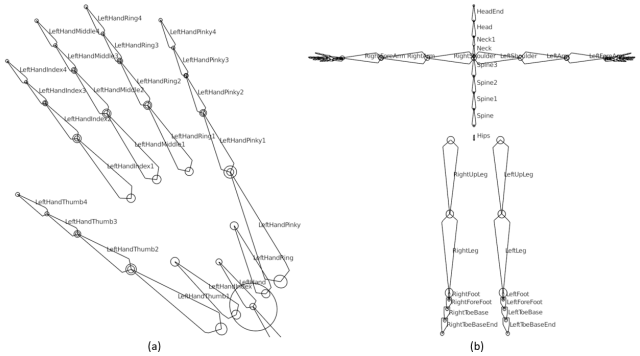


Fig. 5. Joint names for body and hands. The representative human skeleton has 48 hand joints (a) and 27 body joints (b).

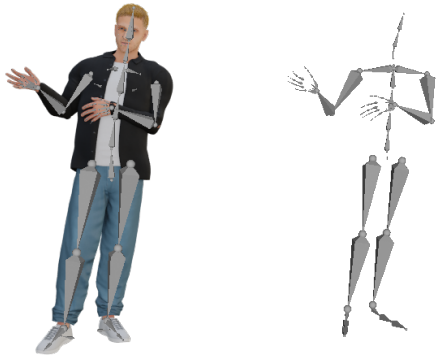


Fig. 6. Motion retargeting example: motion-driving skeleton and retargeted result.

vates one part of the face (*e.g.*, mouth area, eyes, eye-brows) at a time considering the human facial anatomy.

All avatars used for demonstration were built by a Blender tool called HumanGeneratorv3[2]. We created avatars for 8 speakers. Furthermore, we also processed motion retargeting on the body bone animation. Thanks for giving the exception of license from the HumanGenerator team. These facial mesh data will be released together to better visualize the facial data recordings as an asset of the BEAT dataset.

Fig. 7. The names for FACs (Facial Action Coding system) based blendshapes.

D Additional Discussions for SRGR, FGD and BeatAlign.

In the main paper, we demonstrate the SRGR is closer to human perception in the terms of diversity and attractiveness in comparison to the L1 Diversity in [3], considering the score distribution for each group of gesture clips. Here, we list the sum of scores distribution from all gesture clips in Figure 8. The experimental results are shown in Figure 8 (left), which implies that there is a strong correlation between the attractiveness of a gesture and its diversity. More importantly, Figure 8 (right) shows that SRGR is closer to the human perception

[2] https://www.humgen3d.com/

in evaluating the diversity and attractiveness of gesture than the equal weight sum of L1 distance.

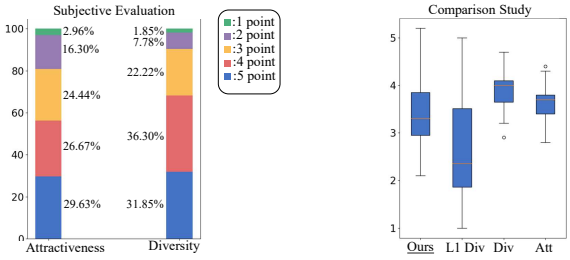


Fig. 8. User Study Results for SRGR. *Left*: The proportional histogram shows the distribution of the score of 5-point Likert scale. *Right*: The scores of comparison study illustrate that SRGR is more akin to subjective human perception in term of diversity and attractiveness when evaluating a gesture.

In addition, we also investigate two other metrics: Frechet Gesture Distance (FGD) [6] and BeatAlign [4]. We observed that they have few limitations for gesture synthesis evaluation. The calculation of FGD mainly depends on the gesture feature representation, but there is no common or well-defined gesture feature representation standard, due to different number of joints and different duration of analyzed segments. Besides, we found that some synthesized gesture sequences in the results received relatively constant FGD scores, however, there are obvious jitters that occurred in these gesture sequences. Although this problem might be solved by evaluating jointly with BeatAlign, it still suggests a better detection of physical correctness, or exploring a more generalized gesture feature extraction network.

The correctness of BeatAlign [4] was verified on dance generation. Therefore, we propose to verify the feasibility and error rate of it on conversational gesture generation. Unlike dance generation, we extract the RMS onset of the audio as the audio beat, and for the gestures we take the local minimum of the velocity. The experimental results show that: i) for a random sample of 300 gestures clips, there is a 6% higher score than GT; ii) for 100 gestures clips with 0.1 second steps and five seconds of panning back and forth, there is an average precision of 83%; iii) Single directional evaluation has higher precision than the bi-directional evaluation (71%), and non-exponential evaluation (59%). Thus, although BeatAlign can be used as a metric to evaluate audio-gesture synchrony in conversational gesture generation, there is still room for improvement.

We also list formulas of L1 Div., FGD, and BeatAlign bellow for reference: L1 diversity, which is the equal weight sum of L1 distance from different N clips, as

$$L1Div. = \frac{1}{2N(N-1)} \sum_{t=1}^{N} \sum_{j=1}^{N} \left\| p_t^i - \hat{p}_t^j \right\|_1, \tag{1}$$

FGD[6] is the FID calculated by a pretrained gesture encoder, as

$$\mathrm{FGD}(\mathbf{m}, \hat{\mathbf{m}}) = \|\mu_r - \mu_g\|^2 + \mathrm{Tr}\left(\Sigma_r + \Sigma_g - 2\left(\Sigma_r \Sigma_g\right)^{1/2}\right), \tag{2}$$

where μ_r and Σ_r are the first and second moments of the latent features distribution z_r of real human gestures $\mathbf{m}$, and μ_g and Σ_g are the first and second moment of the latent features distribution z_g of generated gestures $\hat{\mathbf{m}}$. BeatAlign [4] is calculated as

$$\mathrm{BeatAlign} = \frac{1}{G}\sum_{b_G \in G} \exp\left(-\frac{\min_{b_A \in A}\|b_G - b_A\|^2}{2\sigma^2}\right), \tag{3}$$

Where G, A is the set of gesture beat and audio beat, respectively. σ is adjusted based on fps and we set 0.3 in our paper.

E More Subjective Results and Videos.

The subjective results of generated gestures are shown in Figure 9. More video results and data are available on the project page. We also provide results on other languages, *e.g.*, Japanese data.

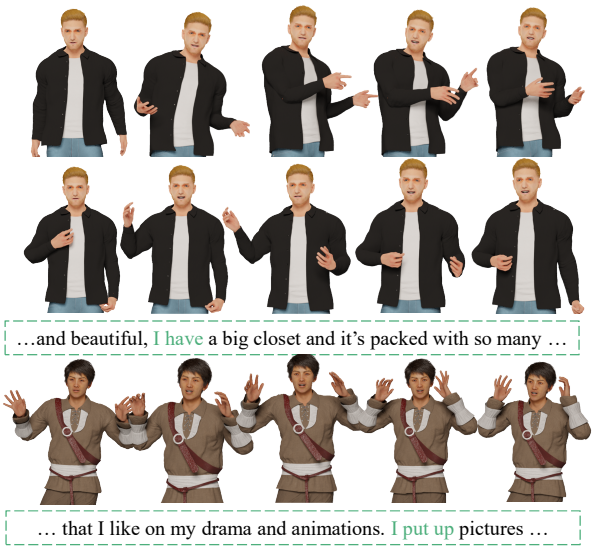


Fig. 9. Results Visualization. Ground truth (top) and generated results with neutral (middle) and fear (down) emotions.

F Details of baseline training

Currently, we do not split the dataset based on speaker, *i.e.*, some speakers only exist in the validation/test data, since the speaker ID is one of the inputs. For each speaker, we use the ratio 10:1:1 for the train/valid/test data splits. For baseline training, we select the best model based on the lowest validation FGD score during training for all baseline models. The final selected epoch is listed in Table 4.

Table 4. Best Epoch for Baselines.

	Seq2Seq	S2G	A2G	MultiContext	Ours (CaMN)
FGD	261.3	256.7	223.8	176.2	123.7
Learning Rate	1e-3	1e-4	1e-4	5e-4	2e-4
Epoch	103	87	271	129	117

References

1. Dutta, A., Zisserman, A.: The VIA annotation software for images, audio and video. In: Proceedings of the 27th ACM International Conference on Multimedia. MM '19, ACM, New York, NY, USA (2019). https://doi.org/10.1145/3343031.3350535, https://doi.org/10.1145/3343031.3350535 1
2. Hornby, A.S., et al.: Oxford advanced learner's dictionary of current English. Oxford University Press, Oxford.[OALDCE] 2, 3
3. 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) 7
4. 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) 8, 9
5. 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) 1
6. 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) 8, 9