1. Supplementary

This supplementary material demonstrates a video to provide better visual and auditory comparisons of the co-speech gesture generation. We also present some details for networks and the semantic prompt gallery.

1.1. SEEG networks

We show details for the network designs in SEEG. All structures of the networks are shown in Fig. 1. Both GRU networks in E_b and E_s are four-layer GRU networks, which are similar to [2].

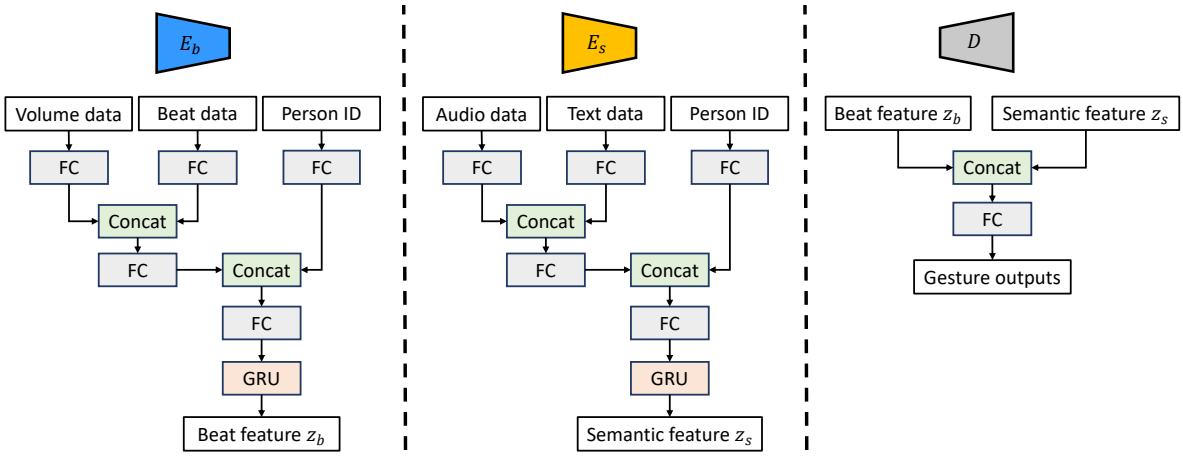

Figure 1. Structures of the networks in SEEG.

1.2. Ablation Study

Effect of Person ID: We use person ID to provide unique style variations for different speakers. The person ID helps the network to learn specific characteristics of speakers. Experiments without person ID only achieve 7.190 and $1.012^{\pm 0.035}$ in FGD and diversity, respectively. The complete SEEG achieves 6.244 and $1.059^{\pm 0.045}$. The smaller FGD and larger diversity of our method indicate the better generation quality. Empirically, person ID provides essential information to converge.

Weights for Loss Functions: We have discussed $\mathcal{L}_{reg}$ and $\mathcal{L}_{adv}$ in Line 469-478, which helps the network to generate corresponding to the ground truth. Meanwhile, $\mathcal{L}_{align}$ focuses on semantic expressiveness, which may introduce differences to the ground truth. We consider an overall loss here, *i.e.*, $\mathcal{L} = \lambda_1\mathcal{L}_{reg} + \lambda_2\mathcal{L}_{adv} + \lambda_3\mathcal{L}_{align}$. When $\lambda_1 = 1, \lambda_2 = 1, \lambda_3 = 2$, SEEG achieves $1.095^{\pm 0.038}$ in diversity. When $\lambda_1 = 1, \lambda_2 = 1, \lambda_3 = 1$, SEED achieves $1.059^{\pm 0.045}$. A larger weight for $\mathcal{L}_{align}$ conducts better diversity and semantic expressiveness. We will present more studies and discussions in the revision.

Besides the SAA metric, we have provided diversity (Tab. 2). The baseline achieves $0.759^{\pm 0.029}$, and ours is $1.059^{\pm 0.045}$. SEEG shows better performances in diversity, revealing the diverse generations and better semantic expressiveness of SEEG.

Third-party Evaluation: We train a new prompter network using gesture data from speaker Ellen in [1]. The baseline method achieves 15.651, and SEEG achieves 20.948 under the average SAA metric. This additional evaluation proves the effectiveness of our proposed SEEG.

1.3. Details for Semantic Prompt Gallery

We present all details of the semantic prompt gallery as in Table 1. We uniformly sample five frames from the gesture sequences in the gallery and visualize them in the table. Besides, all words used to describe the corresponding category are also involved in Table 1.

1.4. Discussions for Limitation and Impact

Technically, SEEG also has some limitations. SEEG focuses on semantic expressions. It may not produce gestures following the ground truth. Compared with the actual movements of speakers, generated gestures may tend to provide larger responses and be different significantly. Moreover, the vivid gestures may lead to negative societal impacts. SEEG may be utilized to produce virtual humans and generate fake information to stimulate the particular person.

Category	Words	Gestures
Listing	first, second, third, fourth, fifth, last, only, earliest, next, 1st, 2nd, 3rd, 4th, prior, preliminary, again, once, recently, currently,just, least, then, final, one, two	
Emphasize	confident, glad, pleased, very, glorious, indeed, obviously, certain, actually, hardly, really, strongly, always, huge, large, big, greatest, incredible, severe, quite, giant, proud, clearly, great, extreme	
Deictics	this, they, that, these, those, it, its, ones, one, which, where, whose, who, what, when, I, my, your, you, she, her, he, his, our, ours	
Negative	injury, hurt, harm, worry, anger, hate, fear, insult, break, terrible, horrible, awful, ugly, wrong, no, dont, rude, unfortunate, dead, risk, ridiculous, weird, broken, bad, dangerous	
Positive	good, excellent, better, nice, lovely, wonderful, yeah, hey, care, favor, joy, live, happy, admire, super, pleasing, honest, pleasant, wish, smile, kiss, hug, yes, healthy, favour	

Table 1. Presentation for the words and gestures in the gallery. In each class, all five gesture sequences are presented. Every raw of the gestures are uniformly and sequentially sampled from the same sequence.

References

[1] Shiry Ginosar, Amir Bar, Gefen Kohavi, Caroline Chan, Andrew Owens, and Jitendra Malik. Learning individual styles of conversational gesture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3497–3506, 2019. 1

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