

# Supplementary Material: Emphasizing Semantic Consistency of Salient Posture for Speech-Driven Gesture Generation

Anonymous Authors

## 1 OVERVIEW

This supplementary material mainly includes:

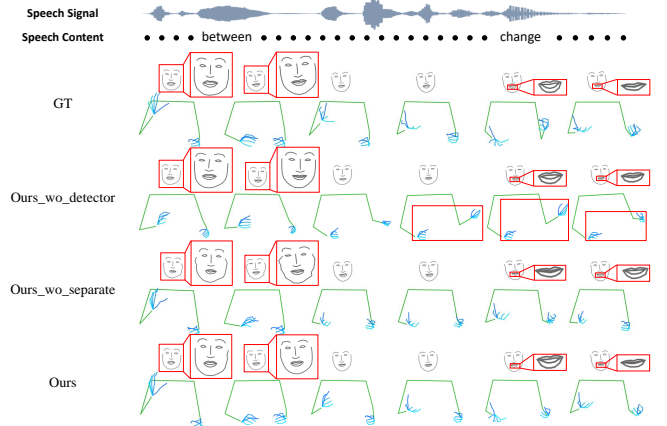
- More ablation study analysis in Section 2.
- Implementation details about the Pose-Sync Distance Metric in Section 3.
- More visual comparisons in Section 4.

We also provide a demo video to show the video results and the code will be publicly released.

## 2 ABLATION STUDY

**Ablation visualization.** In Section 4.5 of the main paper, we conduct ablation studies to justify the contribution of each module to our overall framework and report the quantitative results in Table 3 of the manuscript. Here, we further show the qualitative visualization results as depicted in Figure 1. From top to bottom row: the input speech signal, the speech content, and the gesture sequences from ground truth, our method without salient posture detector, our method without separate synthesis, and our full method. Without the integration of salient posture detector, the crucial gestures related to the semantics of audio fail to be captured (see the solid red box in the second row of Figure 1). While with the salient posture detector, our model generates gestures with salient postures related to the strong semantics, which indicates the salient gesture detection can effectively enforce the semantic consistency between audio and gesture, and is crucial for generating vivid and realistic co-speech gestures. In addition, as shown in the third row of gesture sequences, without the separate synthesis branch, the generated facial expressions and lip motions suffer from poor naturalness and terrible synchrony, which demonstrates the effectiveness of this proposed structure.

**Ablation of top-k frames.** During the training stage of the weakly-supervised salient posture detector, we use the average saliency score of  $top-k$  frames as the final saliency prediction for the video sequence and make it close to the video-sequence-level saliency label. Here, to investigate the effect of hyper-parameter  $top-k$  on the performance of network, we conduct an ablation study of the value of  $k$  on the speaker Kubinec from Speech2Gesture dataset [2]. We select four sets of hyper-parameter  $k$  and report the evaluation results in Table 1. We can see that the value of  $k$  has more significant effect on FGD and PSD metrics than BC metric, which demonstrates that the appropriate parameter contributes to generating more realistic gesture sequences while synchronization is more robust to the changes of the value  $k$ . Therefore, we empirically set  $k = 16$  for the best performance of the generation network.



**Figure 1: Illustration of the effectiveness of each module in our proposed method. Detector and separate respectively denote salient posture detector module and separate synthesis branch.**

**Table 1: Ablation study on the value of  $top-k$ .**

Metric	$k=4$	$k=8$	$k=16$	$k=32$
FGD ↓	1.72	0.78	<b>0.46</b>	0.52
BC ↑	0.69	<b>0.73</b>	0.72	0.72
PSD ↓	6.32	5.90	<b>5.69</b>	5.98

## 3 DETAILS OF POSE-SYNC DISTANCE METRIC

SyncNet [1] is originally proposed to evaluate the audio-video synchronization between lip motion and speech. Inspired by SyncNet, we propose Pose-Sync Net to measure the synchronization between body gestures and speech signals. Given the audio signal and corresponding body pose sequence, we randomly sample aligned and unaligned audio-pose pairs and utilize contrastive learning to train the network.

**Pose-Sync Net.** The detailed structure of our PoseSyncNet is shown in Figure 2. In order to visually evaluate the synchrony between body gesture and speech, we transform the pose coordinates into single-channel image. During the training stage, given an audio sequence  $\mathbf{a} = [A_1, \dots, A_T]$  and corresponding pose sequence  $\mathbf{p} = [P_1^I, \dots, P_T^I]$ , we first generate contrastive training data by randomly sampling audio clip  $a_i = [A_i, \dots, A_{i+8}]$  and pose clip  $p_j = [P_j^I, \dots, P_{j+8}^I]$  with binary similarity label  $y \in \{0, 1\}$  between  $a_i$  and  $p_j$ . For positive pairs, audio clips and pose clips are aligned ( $i = j, y = 1$ ), and for negative pairs, audio clips and pose clips are unaligned ( $i \neq j, y = 0$ ). Then  $a_i$  is transformed into the MFCC feature and processed by an audio encoder to obtain audio feature

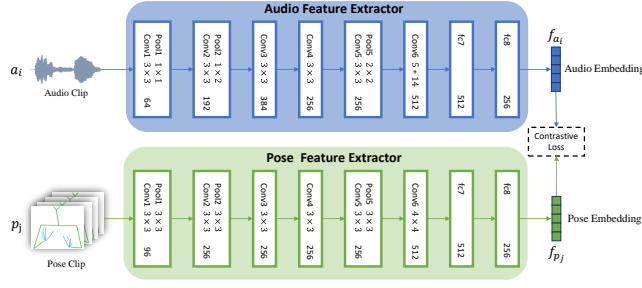


Figure 2: The detailed structure of proposed Pose-Sync Net.

$f_{a_i}$ . After that,  $p_i$  is processed by a pose encoder to obtain pose embedding  $f_{p_j}$ . Then the contrastive loss for training is computed as:

$$L_{PSN} = \frac{1}{2N} \sum_{k=1}^N y \cdot d^2 + (1 - y) \cdot \max(\phi - d, 0)^2, \quad (1)$$

$$d = \|f_{k,a_i} - f_{k,p_j}\|_2, \quad (2)$$

where  $N$  is the batch size and  $\phi$  is a constant,  $d$  is the  $L_2$  distance between audio feature  $f_{k,a_i}$  and pose feature  $f_{k,p_j}$  with  $k$  denoting their index in batch.

During the evaluation process, we sample aligned audio clip and pose clip from the generated pose sequence and use the  $L_2$  distance  $d$  as our Pose Sync Distance (PSD) metric.

## 4 MORE VISUAL COMPARISONS

In this section, we show more visual comparison results with state-of-the-art methods on both datasets in Figure 3 and Figure 4. For the TED Expressive dataset [5], we select one representative case and show the keyframes of all the methods in Figure 3. Compared methods tend to generate unnatural poses and produce unreliable and stiff results. The proposed method can generate realistic human-like poses without resulting in mean poses that are slow and rigid.

For the Speech2Gesture [2] dataset, we select four cases containing strong semantic information from test set to demonstrate the effectiveness of our method, and compare the generated gesture sequences of our method with state-of-the-art methods. The results show that only our method is capable of generating salient postures corresponding to the strong semantic information of the input speech signal. As shown in Figure 4, (1) in the first case, when the speaker says the word *remove*, his left hand makes an inward and then outward swing to indicate the semantics of *remove*. SEEG [4] only learns the outward movement while our method learns the whole salient gesture. (2) In the second case, the speaker brings the hands together underneath, then holds the left hand still and raises the right hand to express the meaning of the word *expand*. Audio2Body [7] learns the wrong direction of hand movement. MoGlow [3] generates the overall gesture appearance but with little movements. SEEG [4] slightly learns the trend of hand movement but the generated hands are less realistic. Our method successfully generates this salient posture with better hand shape and lip synchronization. (3) In the third case, the speaker opens hands and raises the left and right hands alternately to express the meaning of the word *change*. Gestures generated by Audio2Body [7] and MoGlow [3] only raise the left hand while gestures from SDT [6]

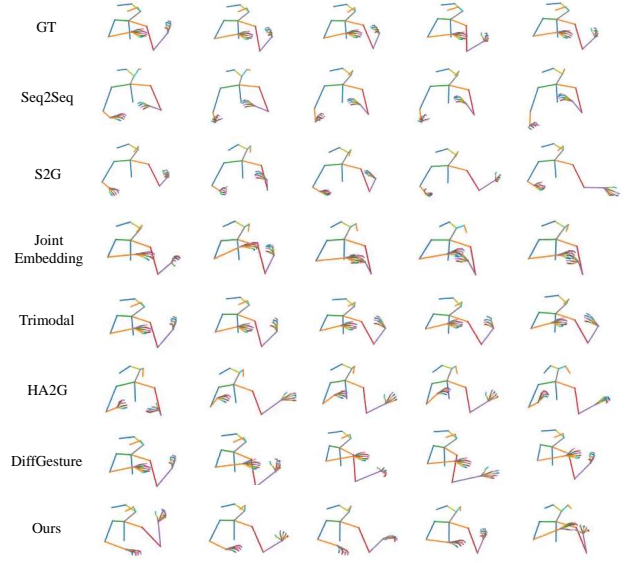


Figure 3: The visual comparisons with state-of-the-art methods on the case sequence of TED Expressive dataset [5].

only raise the right hand. Gestures from S2G [2] mostly keep still with little movement. SEEG [4] generates exceptionally small hands and distorted faces, which seriously affect the realism of results. Our method generates natural and realistic results with better synchronization. (4) In the last case, the speaker obviously raises his right hand and puts down his left hand when he says the phrase *iso-electronic configuration*. Audio2Body [7], S2G [2], and MoGlow [3] generate gestures of similar appearance with a small range of motion and unnatural hands. The template of SDT [6] restricts it from generating poses with large variation, thus it fails to learn strong semantic gestures. SEEG [4] generates the coarse pose appearance, but the generated right hand is sagging while the one of ground truth is flat. Compared with these methods, our approach succeeds in learning the salient posture and generates more realistic results.

## REFERENCES

- [1] Joon Son Chung and Andrew Zisserman. 2016. Out of Time: Automated Lip Sync in the Wild. *asian conference on computer vision* (2016).
- [2] Shiry Ginosar, Amir Bar, Gefen Kohavi, Caroline Chan, Andrew Owens, and Jitendra Malik. 2019. Learning individual styles of conversational gesture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3497–3506.
- [3] Gustav Eje Henter, Simon Alexanderson, and Jonas Beskow. 2020. Moglow: Probabilistic and controllable motion synthesis using normalising flows. *ACM Transactions on Graphics (TOG)* 39, 6 (2020), 1–14.
- [4] 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*. 10473–10482.
- [5] Xian Liu, Qianyi Wu, Hang Zhou, Yinghao Xu, Rui Qian, Xinyi Lin, Xiaowei Zhou, Wayne Wu, Bo Dai, and Bolei Zhou. 2022. Learning Hierarchical Cross-Modal Association for Co-Speech Gesture Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10462–10472.
- [6] Shenhan Qian, Zhi Tu, Yihao Zhi, Wen Liu, and Shenghua Gao. 2021. Speech drives templates: Co-speech gesture synthesis with learned templates. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 11077–11086.
- [7] Eli Shlizerman, Lucio Dery, Hayden Schoen, and Ira Kemelmacher-Shlizerman. 2018. Audio to body dynamics. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 7574–7583.

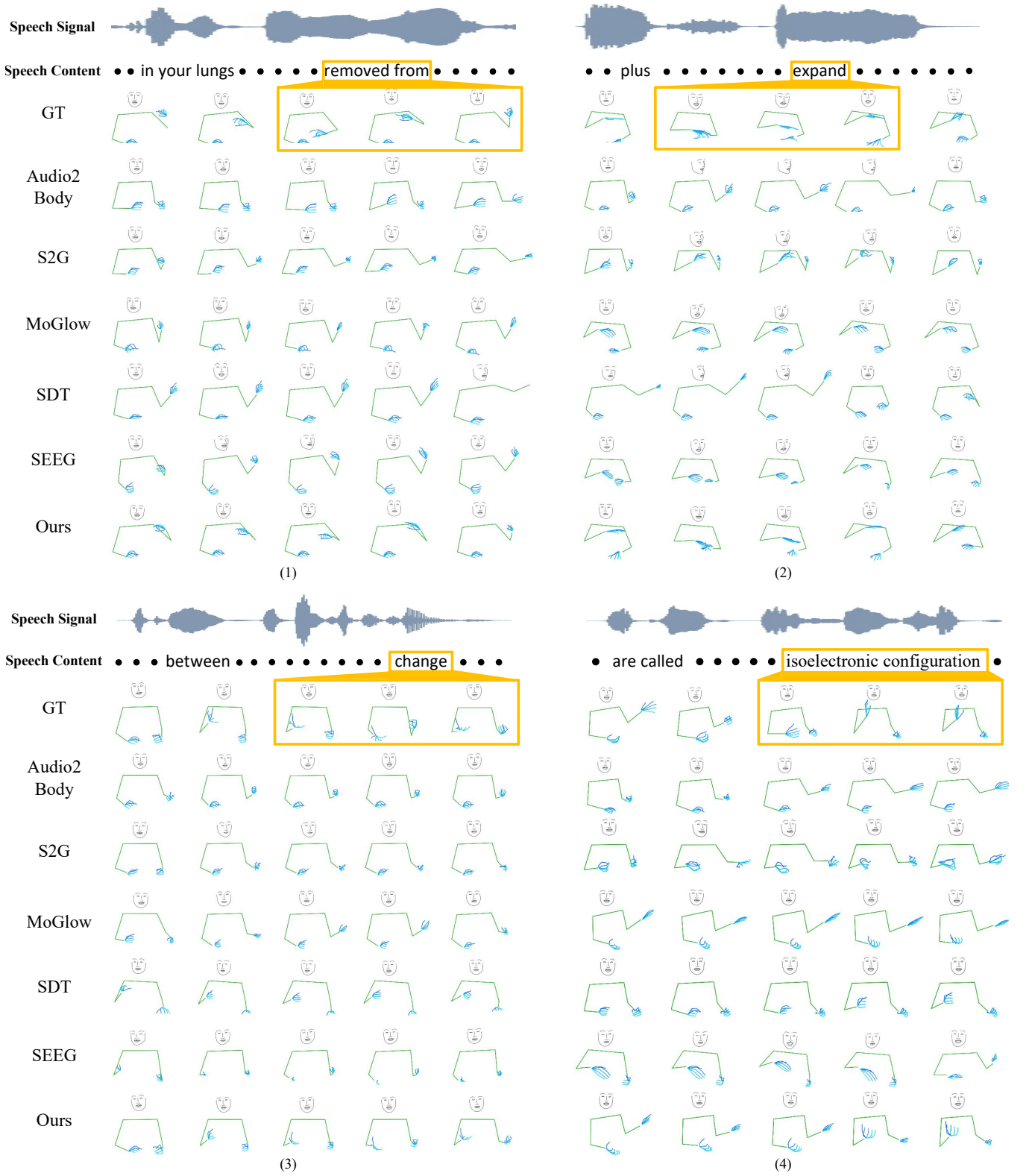


Figure 4: The visual comparisons with state-of-the-art methods on four case sequences of Speech2Gesture [2] dataset. For clarity, we show the key frames of the generated gestures of all methods given the speech signal. Our method can synthesize more natural and realistic gestures with better synchrony than other methods.