LivelySpeaker: Towards Semantic-Aware Co-Speech Gesture Generation

Yihao Zhi[1,*] Xiaodong Cun[2,*] Xuelin Chen[2] Xi Shen[3]
Wen Guo[4] Shaoli Huang[2] Shenghua Gao[1,5,6,†]

[1]ShanghaiTech University [2]Tencent AI Lab [3]Intellindust [4]INRIA
[5]Shanghai Engineering Research Center of Intelligent Vision and Imaging
[6]Shanghai Engineering Research Center of Energy Efficient and Custom AI IC

https://github.com/zyhbili/LivelySpeaker

In the supplementary material, we provide a **supplementary video** to show:

- The pipeline of the whole model as in the main paper.
- The comparison with baselines (Sec. 1).
- The effectiveness of each component in our framework (Sec. 2).
- The applications of interpolating poses between different modalities (Sec. 3).
- The applications of semantic motion generation via new text prompt (Sec. 4).

We also give some explanations aligned with the video and list below.

1. Comparisons with baselines

We show the results comparing to all baselines [2, 4, 5, 8]. On the TED [2] dataset, it is noticeable that HA2G [5], Speech2Gesture [2], and Trimodal [8] generate gestures with rhythmic patterns but lack semantic meaning. Meanwhile, there exists unnatural arm twitching in HA2G [5]. In contrast, our full pipeline outperforms these baselines by excelling in both semantics and rhythm. On the Beat dataset [4], our method shows better visual performance than the state-of-the-art CaMN [4] that utilizes more modalities. Besides, our approach exhibits greater diversity.

2. Individual gestures from each generator

We present the generation results of our individual generators. As shown in our video, regarding semantics, semantic-aware generator (SAG) yields open arms for '*many many*', whereas our rhythm-aware generator (RAG) merely produces waving hands in response to the audio input. However, when the human voice is finished, the output of SAG continues moving while those of RAG become still. Thus, SAG is capable of producing gestures with good content but poor rhythm. As for rhythm, our RAG can generate rhythmic-aware results with little semantics.

3. Application: Interpolating poses between two modalities

To combine the merits of SAG and RAG, we employ our RAG as a beat empowerment module, allowing for editing given motion by adding K steps noise first and then, denoising it through a trained gesture diffusion model. By adjusting the value of K, we can control the semantics and prosody of gestures as well. Here we exhibit the results under adding different noise steps $K \in \{10, 20, 50, 100\}$. The leftmost one ($K = 0$) is the semantic-aware gesture generated from SAG. On its right, we list the edited version of it under different inversion steps. We can observe that when K is small (~ 20), it exhibits both good semantics and rhythm. As the value of K increases, the rhythm-aware gestures dominate the result. However, if the value of K exceeds the threshold (*e.g.* $K > 50$), the semantic gestures will influence a little.

4. Application: Semantic gesture generation via new text prompt

In our SAG, the motion space is well aligned with the text space of CLIP [7]. Inspired by recent advancements in image editing [3, 6] through the prompts, we can easily modify and customize the motion in the same manner. As shown in the supplementary video, we present the results directly obtained from SAG, along with the edited outcome achieved by

*Equal contribution
†Corresponding author.

incorporating specific prompts. For instance, we can roughly manipulate the height and range of gestures by providing the prompts such as "*high*", "*down*", "*many*", *etc*. We also show an example that when we add the prompt like "*in a confirm attitude*", it results in a firm waving down motion. We can also observe similar results on the BEAT [4] dataset. Please view our supplementary video for more details.

5. Inference speed

Our two-stage system, which particularly incorporates a diffusion model, is inherently slower during inference time when compared with GAN-based methods.

For speed comparison, we generate a long sequence consisting of 12k frames (∼800s) using each method and report their running time in Table. 1. The speed of SAG-only is comparable to previous methods while incorporating the diffusion process ($K = 20$ steps) into our full system increases the running time. Nonetheless, there are various advanced sampling techniques for diffusion models that can be suitable for our method. We believe that future, more advanced sampling techniques can benefit our full pipeline.

Methods	S2G	TriModal	HA2G	SAG	Ours Full
Time(s)	2.9	3.1	10.8	5.4	42.6

Table 1. We conduct the experiment on a single RTX 3090.

6. Ablation studies on RAG

The use of MLPs is inspired by recent work on motion prediction [1]. The 1x1 Conv is a linear layer on the temporal axis. Each MLP block adopts the skip connections, the output from the previous MLP layer is added to the output of its subsequent MLP layer. We choose the hyper-parameter experimentally. Here we present detailed ablation studies on TED in Table. 2, where our choice produces the best FGD and Diversity score.

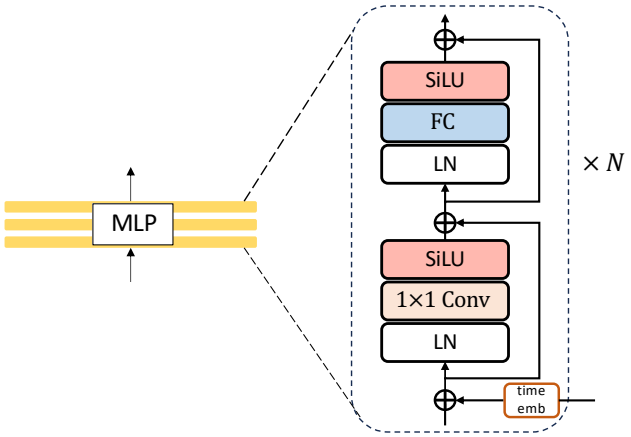

Figure 1. Details of the MLP block.

#	Act.	FGD↓	BC ↑	Diversity↑
4	SiLU	2.152	0.656	107.988
4	ReLU	3.956	0.683	106.581
4	LReLU	5.847	0.682	105.668
4	LReLU†	6.392	0.695	104.497
2	SiLU	8.243	0.689	106.115
6	SiLU	3.047	0.623	104.880
8	SiLU	4.184	0.655	104.876

Table 2. MLP architecture ablation. LReLU and LRELU† represent the LeakyRELU with the scope of 0.1 and 0.2, respectively. # represents the layer of MLP in the backbone.

References

[1] Yuming Du, Robin Kips, Albert Pumarola, Sebastian Starke, Ali Thabet, and Artsiom Sanakoyeu. Avatars grow legs: Generating smooth human motion from sparse tracking inputs with diffusion model. In *CVPR*, 2023. 2

[2] S. Ginosar, A. Bar, G. Kohavi, C. Chan, A. Owens, and J. Malik. Learning individual styles of conversational gesture. In *CVPR*. IEEE, June 2019. 1

[3] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022. 1

[4] Haiyang Liu, Zihao Zhu, Naoya Iwamoto, Yichen Peng, Zhengqing Li, You Zhou, Elif Bozkurt, and Bo Zheng. Beat: A large-scale semantic and emotional multi-modal dataset for conversational gestures synthesis. In *ECCV*, pages 612–630. Springer, 2022. 1, 2

[5] Xian Liu, Qianyi Wu, Hang Zhou, Yinghao Xu, Rui Qian, Xinyi Lin, Xiaowei Zhou, Wayne Wu, Bo Dai, and Bolei Zhou. Learning hierarchical cross-modal association for co-speech gesture generation. In *CVPR*, pages 10462–10472, 2022. 1

[6] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. *arXiv preprint arXiv:2211.09794*, 2022. 1

[7] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, 2021. 1

[8] 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