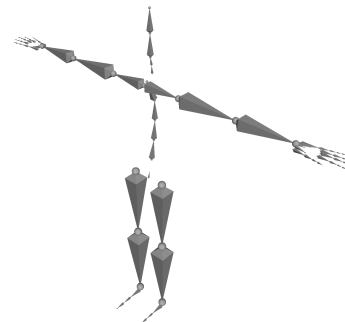


Anonymous Authors

- (1) An in-depth examination of the human gesture data format adopted in our research, with specifics presented in Section 1
- (2) A detailed description of the multi-modal feature processing techniques that underpin our model, discussed in Section 2.
- (3) A thorough analysis of the ablation studies that underscore the efficacy of the individual elements within our framework, as detailed in Section 3.
- (4) A comparison of the model complexity and computational costs, in terms of parameters and FLOPs, which is provided in Section 4.

Emotion: The BEAT dataset’s eight emotions are each encoded as a one-hot vector and transformed by a linear layer into the emotion feature $f_e \in \mathcal{R}^{12}$. The emotion encoder E_e converts these vectors



(a)

- LefHandThumb4
- LefHandThumb3
- LefHandThumb2
- LefHandThumb1
- LefHandIndex4
- LefHandIndex3
- LefHandIndex2
- LefHandIndex1
- LefHandRing4
- LefHandRing3
- LefHandRing2
- LefHandRing1
- LefHandMiddle4
- LefHandMiddle3
- LefHandMiddle2
- LefHandMiddle1
- LefHandPinky4
- LefHandPinky3
- LefHandPinky2
- LefHandPinky1

(b)

- HeadEnd
- Head
- Neck1
- Neck2
- Spine4
- Spine3
- Spine2
- Spine1
- Spine
- Hips
- Knees
- Ankles
- Feet
- RightFoot
- LeftFoot
- RightForefoot
- LeftForefoot
- RightBallEnd
- LeftBallEnd
- RightHeelEnd
- LeftHeelEnd

Others: To enhance the model’s robustness, the style feature f_s and emotion feature f_e are subjected to random masking based on a Bernoulli distribution during training, simulating the variability inherent in real-world scenarios.

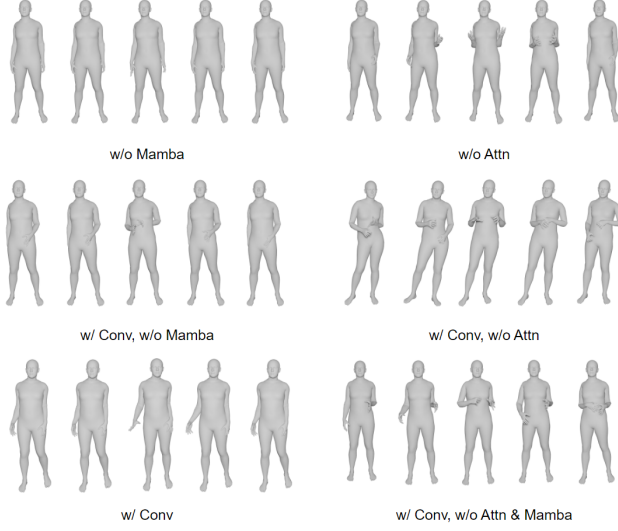


Figure 3: Visualization of different MambaAttn block designs, with the speech transcript: "... when you have to work Monday through Friday the whole week, you are very tired ...", consistent with previous figures.

3 ADDITIONAL DETAILS OF ABLATION STUDIES

In this section, we present further visualizations of the ablation studies conducted on the MambaAttn block design and feature fusion modules. Figure 3 displays the various configurations of the MambaAttn block. We experimented with adding a convolutional layer before self-attention, arranging the block sequence to include a convolutional layer with a kernel size of 3, followed by a self-attention layer and a Mamba layer, with layer normalization at both the beginning and end. This design is predicated on the hypothesis that convolution can capture local information, self-attention can grasp global context, and the Mamba layer can provide sequential modeling.

Figure 4 visualizes the results from our ablation study on feature fusion modules, showcasing the original DSG+ input module, SA fusion module, SEA fusion module, and SEAD-basic fusion module.

Additionally, we illustrate the SEA feature fusion module. Compared to the SEA fusion module, the SA feature fusion module does not incorporate emotion as a condition, highlighting the impact of including emotional context in the fusion process.

These visualizations provide a clear comparison of the different design choices within our MambaAttn block and feature fusion modules, underscoring the importance of each component in enhancing the quality of co-speech gesture generation.

4 PARAMS AND FLOPS

Our MambaGesture framework, as detailed in Table 1, demonstrates a notable advancement over existing state-of-the-art methods in co-speech gesture generation. Our model exhibits a higher parameter count at 6.443 million, compared to CaMN's 0.303 million, MDM's 3.691 million, and DSG+'s 3.629 million. This increase reflects the

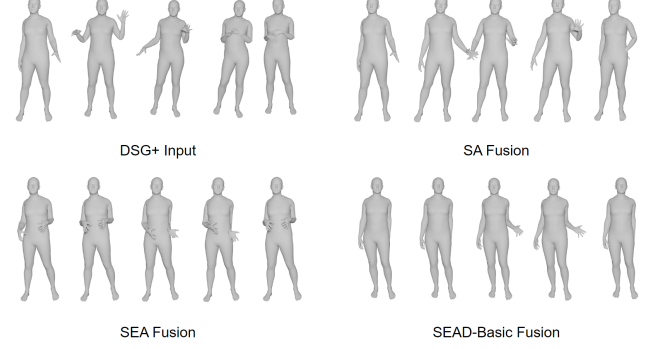


Figure 4: Comparative visualization of different feature fusion modules, using the speech transcript: "... when you have to work Monday through Friday the whole week, you are very tired ...", for consistency with related visualizations.

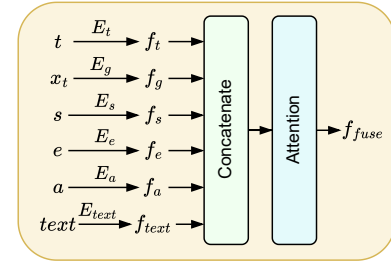


Figure 5: The SEA feature fusion module, demonstrating the integration of style and emotion into the feature fusion process.

complexity of our architecture, which incorporates advanced multi-modal data processing components and the innovative MambaAttn block.

In terms of computational efficiency, our method requires 33.229 billion FLOPs, which is more efficient than CaMN's 51.377 billion and substantially lower than MDM's 56.952 billion. While DSG+ operates at the lowest computational cost of 12.416 billion FLOPs, our model's enhanced performance justifies the additional computational expense.

Focusing on the Mamba-only variant of our architecture, which excludes the attention mechanism from MambaAttn block in the denoising process, we observe a significant reduction in parameters to 3.131 million and FLOPs to 22.183 billion. Despite this simplification, the Mamba-only model achieves commendable results, with a competitive FGD Score, high Diversity Score, and the highest Beat-Align score among all methods. This underscores the efficacy of the Mamba model in efficiently fusing multi-modal features while maintaining high performance.

The balance between computational cost and performance is a critical aspect of model design. Our MambaGesture framework demonstrates that the trade-off is well warranted, with the increased Params and FLOPs contributing to the state-of-the-art performance in co-speech gesture generation, as evidenced by our extensive experimental evaluation.

Table 1: Comparison of params and FLOPs across different methods, highlighting the effectiveness of our MambaGesture framework. The best is bold, and the second is underlined.

Method	FGD Score↓	Diversity Score↑	L1Div Score↑	SRGR Score↑	BeatAlign↑	Params (M)	FLOPs (G)
GT	-	395.20	850.51	-	0.893	-	-
CaMN	65.74	277.06	587.12	0.241	0.819	0.303	51.377
MDM	106.56	331.53	<u>1001.52</u>	<u>0.229</u>	0.810	3.691	56.952
DSG+	103.15	352.31	789.83	0.238	0.841	3.629	12.416
Ours (Mamba-only)	<u>46.58</u>	<u>358.49</u>	864.19	0.238	0.867	<u>3.131</u>	<u>22.183</u>
Ours	22.11	434.94	1128.79	0.237	<u>0.853</u>	6.443	33.229

The data presented in the table illustrates the superior performance of our full MambaGesture framework, particularly in achieving the lowest FGD Score and the highest scores in Diversity and L1 Diversity, indicating a significant improvement in the quality and variation of generated gestures. Our Mamba-only variant also shows exceptional performance, especially in BeatAlign, confirming the effectiveness of Mamba’s sequence modeling in aligning gestures with audio beats. The trade-off between the number of parameters and computational cost is evident, yet the gains in gesture generation quality affirm the value of our approach.

REFERENCES

[1] Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Jian Wu, Michael Zeng, and Furu Wei. 2021. WavLM: Large-Scale Self-Supervised Pre-training for Full Stack Speech Processing. *arXiv preprint arXiv:2110.13900* (2021).

[2] Haiyang Liu, Zihao Zhu, Naoya Iwamoto, Yichen Peng, Zhengqing Li, You Zhou, Elif Bozkurt, and Bo Zheng. 2022. Beat: A large-scale semantic and emotional multi-modal dataset for conversational gestures synthesis. In *European conference on computer vision*.

[3] Tomas Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhersch, and Armand Joulin. 2018. Advances in Pre-Training Distributed Word Representations. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

[4] Sicheng Yang, Haiwei Xue, Zhensong Zhang, Minglei Li, Zhiyong Wu, Xiaofei Wu, Songcen Xu, and Zonghong Dai. 2023. The DiffuseStyleGesture+ entry to the GENE Challenge 2023. In *Proceedings of the 25th International Conference on Multimodal Interaction*.