

Supplemental

Methods	Method Type	FPS
LS3DCG [2]	GAN	8500
DiffGesture [6]	Diffusion (DDPM)	20
TalkSHOW [5]	VQ-VAE	150
DiffSHEG [1]	Diffusion (DDIM)	65
Ours	Diffusion (DDPM)	10

A.1. **Approximate FPS Measures.** We compare between our method LS3DCG, DiffGesture, TalkSHOW and DiffSHEG.

A. Implementation Details

Network Parameters For the HuBERT encoder we use the large model that is fine-tuned on the 960h of Librispeech speech audio from Meta and the raw speech audio is sampled at 16kHz. The HuBERT features from the encoder are of dimension 1024 which we project to a size of 64 using a single linear layer as described in Sec. 3.3.

The face motion has 103 parameters that is projected into the shared latent space of dimension 64. Likewise, the body with 129 parameters is also projected into the shared latent space of dimension 64. This 64 dimensional latent space is projected by the additional linear layer as discussed in Sec. 3.3 to a shared hidden dimension of size 512. In the reverse, the outgoing linear projections go from size 512 to 103 and 129 for the face and body respectively. The transformer encoder and transformer decoders are both made of stacks of size 4. The adapter module downsamples with a reduction factor of 8 and has 2 trainable latent tokens.

Training and Inference Parameters For training we use the Adam optimizer, with a learning rate of $5e-4$ and $\beta_1 = 0.9$ and $\beta_2 = 0.999$. We train with a batch size of 512 for 250 epochs on two NVidia A5000 GPUs. The network is trained with classifier-free guidance, using a conditional information drop rate of 0.1. For inference, we set the classifier-free guidance scale at 0.15. For training and inference we use $T = 500$ timesteps and we use a linear variance schedule from $\beta_1 = 1e-4$ to $\beta_T = 0.02$.

B. Performance Comparisons

The focus of our work was not on FPS performance, but rather on combining face and body networks into one for

savings on trainable parameters. We compare with a variety of different method types as shown in Tab. A.1. Here we see that the non-diffusion methods perform well on FPS. In contrast, the diffusion methods have much lower performance. Specifically, the two DDPM methods (ours and DiffGesture) have the lowest performance. Here we would like to note the provided DiffGesture metric is for only the body network, to also generate the face would require another pass of a separate network. Both our network and DiffGesture use 500-step DDPM diffusion, whereas DiffSHEG uses a 25-step DDIM diffusion model. Using DDIM results in significant speed ups by omitting diffusion steps as discussed in [4]. Using DDIM or training a latent consistency model (LCM), such as in [3] can provide significant performance gains in inference time with little quality loss, and adapting them to our network provides exciting future work to further improve the performance of face and body speech motion generation.

References

- [1] Junming Chen, Yunfei Liu, Jianan Wang, Ailing Zeng, Yu Li, and Qifeng Chen. Diffsheg: A diffusion-based approach for real-time speech-driven holistic 3d expression and gesture generation. In *CVPR*, 2024. 1
- [2] Ikhsanul Habibie, Weipeng Xu, Dushyant Mehta, Lingjie Liu, Hans-Peter Seidel, Gerard Pons-Moll, Mohamed Elgharib, and Christian Theobalt. Learning speech-driven 3d conversational gestures from video. In *Proceedings of the 21st ACM International Conference on Intelligent Virtual Agents, IVA '21*, page 101–108, New York, NY, USA, 2021. Association for Computing Machinery. 1
- [3] Simian Luo, Yiqin Tan, Longbo Huang, Jian Li, and Hang Zhao. Latent consistency models: Synthesizing high-resolution images with few-step inference, 2023. 1
- [4] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models, 2022. 1
- [5] Hongwei Yi, Hualin Liang, Yifei Liu, Qiong Cao, Yandong Wen, Timo Bolkart, Dacheng Tao, and Michael J Black. Generating holistic 3d human motion from speech. In *CVPR*, 2023. 1
- [6] Lingting Zhu, Xian Liu, Xuanyu Liu, Rui Qian, Ziwei Liu, and Lequan Yu. Taming diffusion models for audio-driven co-speech gesture generation. In *CVPR*, 2023. 1