Democratizing High-Fidelity Co-Speech Gesture Video Generation

Supplementary Material

This supplementary material consists of four sections. Section 1 introduces the project demo we created. In Section 2, we present a user study to validate the performance of our method. Section 3 evaluates how different classifier-free guidance scales affect our method. In Section 4, we conduct experiments to support the efficiency of our audio-to-skeleton prediction method.

1. Project Demonstration

We have created a project demo to showcase additional video results of our method, which has been submitted as part of the supplementary material.

2. User Study

We conducted a user study comparing visual fidelity, audio-lip synchronization, and motion naturalness between our method and existing approaches. 15 participants assessed 30 randomly selected video clips through pairwise comparisons. Table 1 shows that our method achieves superior performance across all metrics, demonstrating significantly improvements in visual fidelity, audio-lip synchronization, and motion naturalness over other methods.

3. Classifier-Free Guidance Scales

We conduct experiments to evaluate the impact of different classifier-free guidance [1] (CFG) scales on model performance. As shown in Table 2, our method achieves the best overall performance when the CFG scale is set to 3.5. Moreover, we observe that further increasing the CFG scale improves audio-lip synchronization but at the cost of degrading the quality of generated videos.

4. Comparisons in Model Complexity

We compare model complexities between our audio-to-skeleton prediction model and two state-of-the-art human video generation models, namely EchoMimicV2 [2] and StableAnimator [3]. All experiments are conducted with a V100 GPU. As shown in Table 3, our model features significantly fewer parameters and achieves significantly higher outputs. Specifically, they process 19.23, 0.07, and 0.15 samples per second, respectively. These results suggest that using our audio-to-skeleton prediction model as a preliminary step for human video generation is practical, as it occupies only a very small portion of the overall computational cost.

Table 1. User study results.

Methods	Visual Fidelity	Audio-Lip Synchronization	Motion Naturalness
MimicMotion	9.1%	14.0%	14.2%
Ours (MimicMotion)	90.9%	86.0%	85.8%
EchoMimicV2	11.7%	16.9%	15.3%
Ours (EchoMimicV2)	88.3%	83.1%	84.7%
StableAnimator	10.0%	14.4%	16.4%
Ours (StableAnimator)	90.0%	85.6%	83.6%

Table 2. Comparisons in classifier-free guidance scales on our CSG-405 database.

CFG Scales	SSIM↑	PSNR↑	CSIM↑	FID↓	FVD↓	Sync-C↑	Sync-D↓
1	0.65	16.22	0.79	66.89	1028.43	4.37	10.41
3.5	**0.67**	**16.62**	**0.82**	**62.93**	**984.13**	**6.28**	**8.68**
6	0.65	16.00	0.77	67.83	1077.16	5.77	9.27
8.5	0.65	15.79	0.75	63.70	1146.75	5.52	9.42

Table 3. Comparisons in model complexity.

Method	Parameters (Million)	Throughput (samples/s)
Audio-to-Skeleton Prediction Model	228	19.23
StableAnimator	1754.2	0.07
EchoMimicV2	2258	0.15

References

[1] J. Ho, T. Salimans. Classifier-free diffusion guidance. In *NeurIPS Workshop*, 2021. 1

[2] R. Meng, X. Zhang, Y. Li, C. Ma. EchoMimicV2: Towards Striking, Simplified, and Semi-Body Human Animation. In *CVPR*, 2025. 1

[3] S. Tu, Z. Xing, X. Han, Z. Cheng, Q. Dai, C. Luo, Z. Wu. StableAnimator: High-Quality Identity-Preserving Human Image Animation. In *CVPR*, 2025. 1

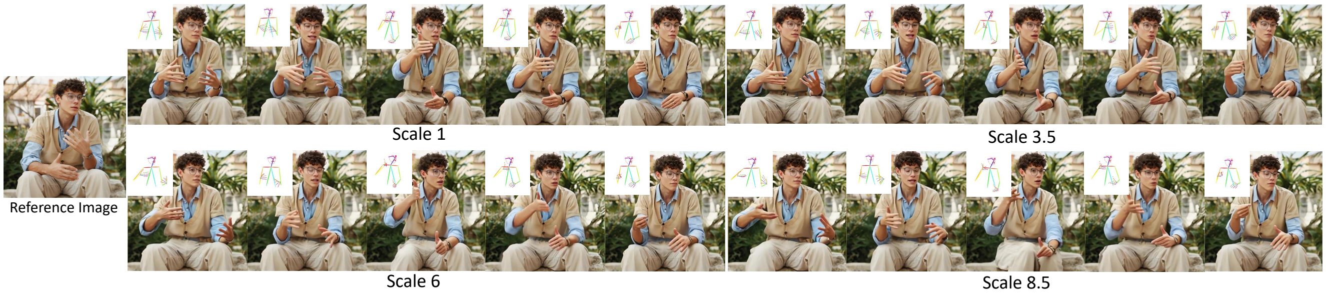

Figure 1. Comparisons of different classifier-free guidance (CFG) scales on our CSG-405 database.