

# Expressive Talking Head Generation with Granular Audio-Visual Control (Supplementary Material)

Borong Liang<sup>1\*</sup> Yan Pan<sup>2,3\*</sup> Zhizhi Guo<sup>1†</sup> Hang Zhou<sup>1†</sup> Zhibin Hong<sup>1</sup>  
Xiaoguang Han<sup>2,3</sup> Junyu Han<sup>1</sup> Jingtuo Liu<sup>1</sup> Errui Ding<sup>1</sup> Jingdong Wang<sup>1</sup>

<sup>1</sup>Department of Computer Vision Technology (VIS), Baidu Inc.,

<sup>2</sup>SSE, CUHK-Shenzhen, <sup>3</sup>FNii, CUHK-Shenzhen

{liangborong, zhouhang09, guozhizhi, hongzhibin, hanjunyu, liujingtuo, dingerrui, wangjingdong}@baidu.com,  
{yanpan@link., hanxiaoguang@cuhk.edu.cn.}

This supplementary material provides a demo video to show the main features about **GC-AVT**, the comparison with previous methods, the ablation study for the main paper.

## 1. Descriptions to the Supplemental Video

The demo video consists of 3 parts, including the feature introduction of Granular facial control, the comparison with previous methods, and the ablation study.

### 1.1. Experimental Settings

In introduction stage of the demo video, our videos are generated based on one identity reference image and driven by a clip of audio with different pose source and expression source. The mouth movement of our generated results should be synced with the audio source. The expressions of our generated results should be controlled by the expression source, while the head poses are controlled by the pose source. In the comparison with previous methods, the results are generated based on one identity reference image with pose, expression, and speech information all coming from a clip of video. The setting for ablation results is the same as the comparison with previous methods.

### 1.2. Comparison with Previous Methods

We compare our method with state of the art audio-driven methods including MakeitTalk [3], Wav2Lip [1] and PC-AVS [2]. Most of the comparing methods do not support granular control such as expression and head pose, therefore it's unfair to set too detailed sources. We keep the rest of the settings the same as ours when re-implementing their results. We use the same pose source and audio source to drive comparing methods as ours. Please note that the sampled cases are very challenging. Ours model generates

much larger body area which makes the task of talking head generation more difficult.

## 2. Ablation Study and Additional Results

### 2.1. Qualitative Results

We show the video results to study the effects of the losses setting and the time-shift operation. For loss setting, we study the effects of VGG loss, VGGFace loss and Contrastive loss. As can be seen in the video, without the perceptual losses such as VGG loss and VGGFace loss, the quality of generated images are obviously poor, and the performance of attribute control is also worse than the results of our complete setting. The speech content driving results are affected when we remove the contrastive loss. The speech driven results are not synced with the driving source. Besides, without the time-shift operation the speech driven results is affected, the movement of lips is not so accurate.

## References

- [1] K R Prajwal, Rudrabha Mukhopadhyay, Vinay P. Nambodiri, and C.V. Jawahar. A lip sync expert is all you need for speech to lip generation in the wild. In *Proceedings of the 28th ACM International Conference on Multimedia (ACMMM)*, 2020. 1
- [2] Hang Zhou, Yasheng Sun, Wayne Wu, Chen Change Loy, Xiaogang Wang, and Ziwei Liu. Pose-controllable talking face generation by implicitly modularized audio-visual representation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 1
- [3] Yang Zhou, Xintong Han, Eli Shechtman, Jose Echevarria, Evangelos Kalogerakis, and Dingzeyu Li. Makeittalk: Speaker-aware talking head animation. *SIGGRAPH ASIA*, 2020. 1

\*Equal Contribution.

†Corresponding author.