

Supplementary Material of VividTalk: One-Shot Audio-Driven Talking Head Generation Based on 3D Hybrid Prior

Xusen Sun¹ Longhao Zhang³ Hao Zhu^{1, ✉} Peng Zhang^{2, ✉} Bang Zhang²
Xinya Ji¹ Kangneng Zhou⁴ Daiheng Gao² Liefeng Bo²
Xun Cao¹
¹Nanjing University ²Alibaba Group ³ByteDance ⁴Nankai University

1. Overview

The supplementary material contains a video and a PDF file. The video shows more dynamic results of our method and the comparison with other state-of-the-art works. In the PDF file, we first show more results to demonstrate the generalization of the proposed method in Section 2. Then introduce the details of network architecture in Section 3.

2. More Experiments

In the main paper, we compared our method with the state-of-the-art methods qualitatively and quantitatively. Here, we conduct several additional experiments to further demonstrate the superiority of the proposed VividTalk.

2.1. Expand to Pseudo Data

With the development of Artificial Intelligence Generated Content, AI can create more and more realistic and useful data, such as images, audio, and videos. Therefore, we hold an opinion that a general talking head generation application can synthesize not only videos based on real data but also pseudo-data, *e.g.* data generated by AI. To this end, we first generate pseudo audio data and facial images with text-to-speech [3] and diffusion model [2], respectively. Then we attempt to drive the facial images according to the input audio with our VividTalk. The visual results are shown in Figure 1, and we strongly recommend watching the dynamic results in the supplementary video. It can be found that our method also works well on pseudo-data, which demonstrates the generalization of VividTalk.

2.2. Expand To Different Language

With benefits from the powerful audio extractor and more decoupled design, our model can also be generalized to other languages, such as Chinese, French, and so on. Here, we generate talking head videos with audio in different languages as input, and the dynamic driven results can be

found in the supplementary video.

2.3. Expand To Long-term Audio Sequence

During training, the audio sequences are clipped into short-term fragments as the input of the network. To generate videos based on long-term audio sequences during the inference stage, we employ a sliding window-based recursive approach to predict motion. Specifically, our model receives a sequence of adjacent past audio features $A = (a_{n-k+1}, \dots, a_{n-1}, a_n)$ with a sliding window in size k , and generates the motion at frame n . In this way, our model can be easily extended to long-term audio sequences without performance degradation. Please refer to the supplementary video for better visual results.

3. Network Architecture

In this section, we introduce the details of the network architecture in our framework.

Global and Local Facial Motion Generator. This network generates blendshape and vertex from the input audio to model the non-rigid facial motion. We use a pre-trained 3D face reconstruction model and audio extractor, of which the detailed architecture can be referred to FaceVerse [4] and wav2vec 2.0 [1]. Then a multi-branch network is proposed to model global and local motion individually. As shown in Figure 2 (a), the network Φ in each branch takes the audio feature A , style embedding z^{style} , and past motion $X^{1 \dots f-1}$ as input, and output the current motion X^f with correspondence dimensions.

Learnable Head Pose Codebook. This is a two phase architecture which consists of an encoder \mathcal{E} , a decoder \mathcal{D} , a mapping network Φ_{map} and a learnable context-rich codebook \mathcal{Z} . As illustrated in Figure 2 (b), (c), the encoder \mathcal{E} and decoder \mathcal{D} are composed of convolutional layers, linear layers, and stacks of multiple Attention Blocks. The mapping network Φ_{map} has a similar structure to Φ in Global

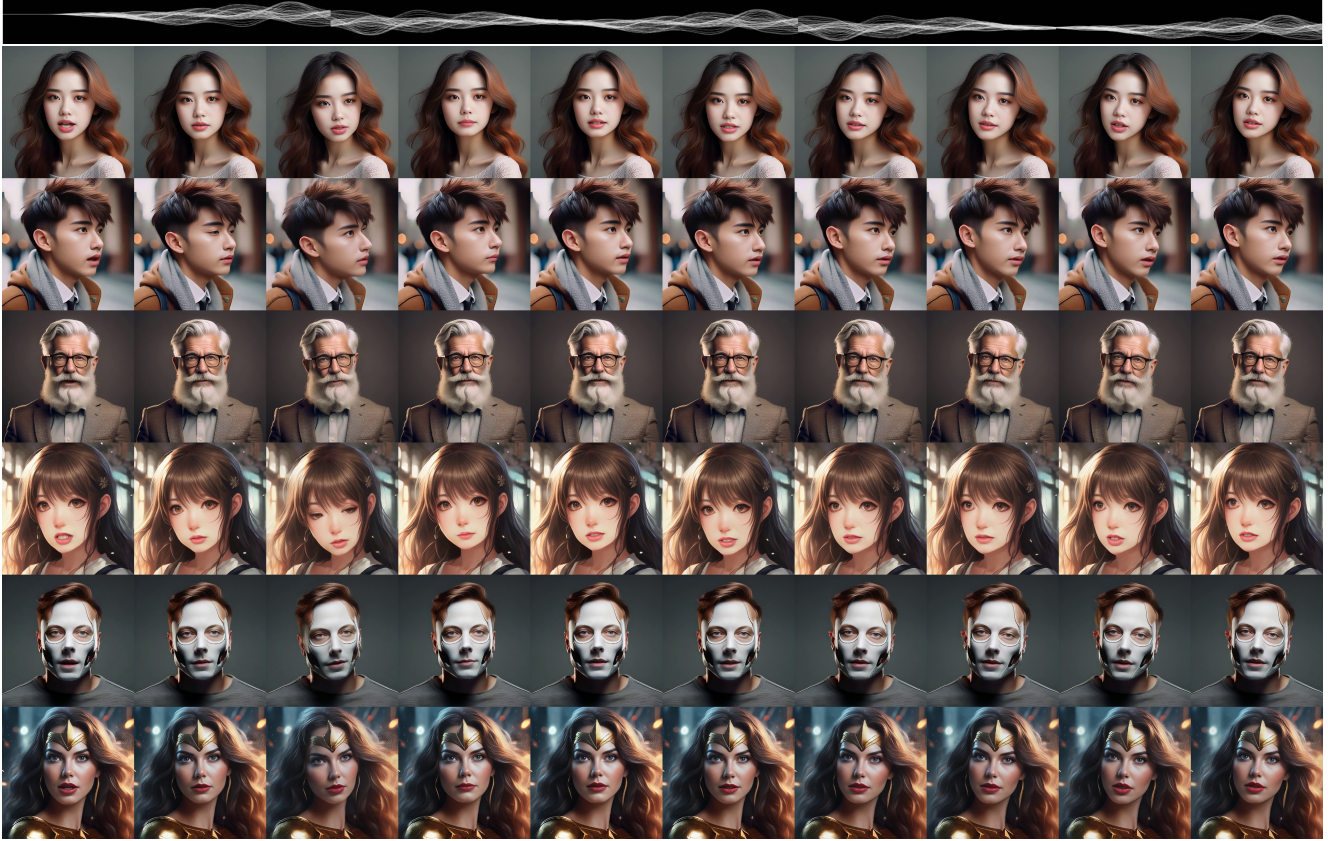


Figure 1. Visual results with pseudo facial images and pseudo audio data as inputs.

and Local Facial Motion Generator, it maps the audio feature A , style embedding s , and initial pose P^0 to the latent space of codebook \mathcal{Z} .

Motion-VAE. As described in the main paper, we first transform the driven meshes obtained in the previous stage into projection texture. Then the motion-vae takes the projection texture as input, which will be mapped to latent space (μ, σ) with 128-dimension. And the reparameterization trick is adopted to sample and synthesize the facial motion map. We also utilize an Hourglass network to enhance the lip motion with the lip-related vertex as input. Finally, the facial motion and the lip motion are concatenated to generate dense motion and occlusion map with 256-dimension. The detailed structure is illustrated in Figure 2 (d).

References

- [1] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460, 2020. [1](#)
- [2] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. [1](#)
- [3] Xu Tan, Jiawei Chen, Haohe Liu, Jian Cong, Chen Zhang, Yanqing Liu, Xi Wang, Yichong Leng, Yuanhao Yi, Lei He, et al. Natralspeech: End-to-end text to speech synthesis with human-level quality. *arXiv preprint arXiv:2205.04421*, 2022. [1](#)
- [4] Lizhen Wang, Zhiyuan Chen, Tao Yu, Chenguang Ma, Liang Li, and Yebin Liu. Faceverse: a fine-grained and detail-controllable 3d face morphable model from a hybrid dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20333–20342, 2022. [1](#)

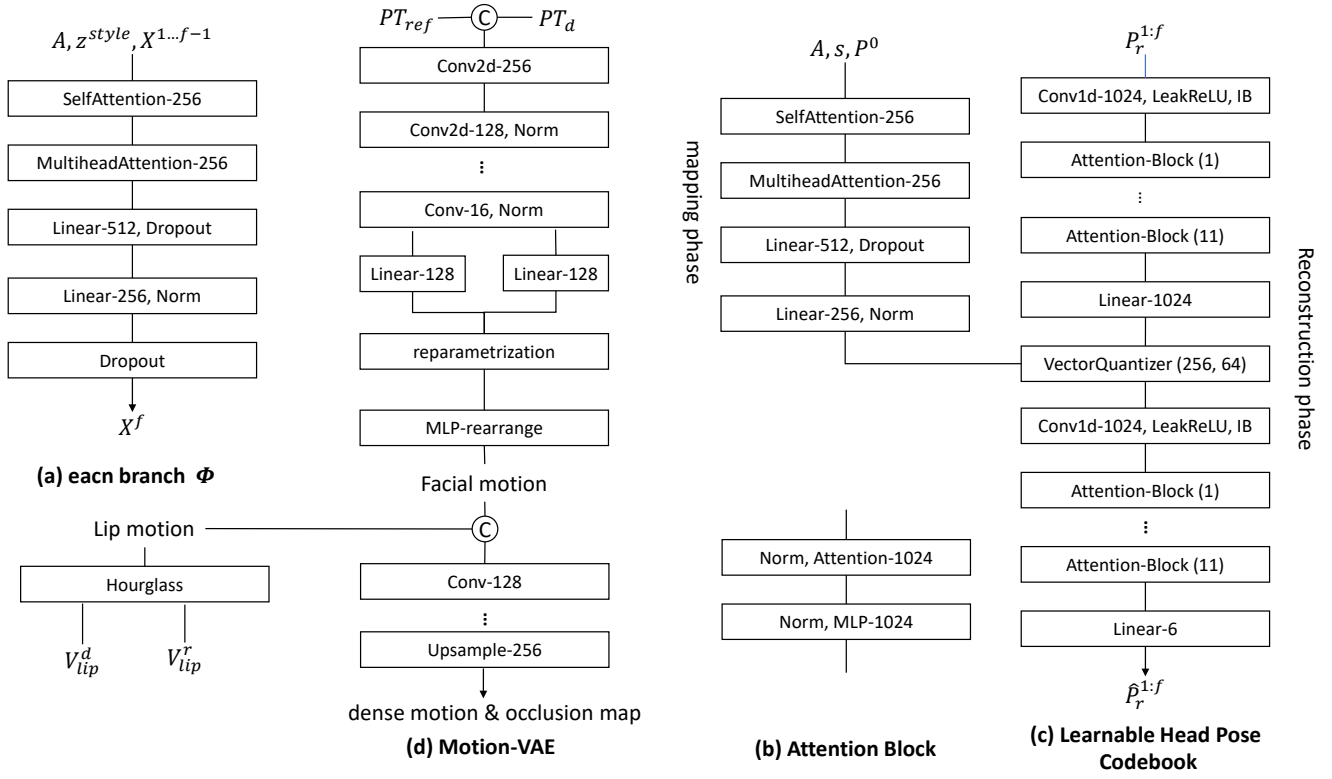


Figure 2. The network architecture of different components in our VividTalk.