Diff2Lip: Audio Conditioned Diffusion Models for Lip-Synchronization (Supplementary Material)

1. Denoising Diffusion Implicit Models (DDIM)

In DDIM formulation [4], the posterior is defined such that one can control its variance schedule $\{\sigma_t \in \mathbb{R}_{\geq 0}\}_{t=1}^{T}$. Hence, in the reverse diffusion process, x_{t-1} can be sampled from x_t using:

$$x_{t-1} = \sqrt{\bar{\alpha}_{t-1}} x_0 + \sqrt{1 - \bar{\alpha}_{t-1} - \sigma_t^2} \left(\frac{x_t - \sqrt{\bar{\alpha}_t} x_0}{\sqrt{1 - \bar{\alpha}_t}} \right) + \sigma_t \xi \quad (1)$$

for $\xi \in \mathcal{N}(0, \mathbf{I})$. If for all t, $\sigma_t = \sqrt{(1 - \bar{\alpha}_{t-1})/(1 - \bar{\alpha}_t)} \sqrt{1 - \bar{\alpha}_t / \bar{\alpha}_{t-1}}$, this formulation becomes the same as standard DDPM [2]. On the other hand, if for all t, $\sigma_t = 0$, then the reverse diffusion process sampling becomes deterministic and simplifies to Eq. (4) from the main text. Further, in this formulation, one can use strided timesteps to make inference faster without training the model again. For example, instead of $t = 1, 2, 3, ..., 1000$, if one uses $t = 40, 80, 120, ..., 1000$, the inference can be done in 25 steps instead of 1000 steps.

2. Qualitative Results

2.1. More Qualitative Figures

Similar to Fig. 6 in the main paper, we show more visual comparisons in Fig. 1. Here we compare our approach with Wav2Lip [3] and PC-AVS [5]. It can be seen in the zoomed-in lip regions that Diff2Lip outperforms other methods both in image quality and identity preservation. Further, we show more examples of reconstruction setting in Fig. 2, 3, 4, and 5 similar to Fig. 4 in main paper. Here we can observe that our method produces the correct lip shape corresponding to the audio, which can be seen when compared with the original video (top row). Finally, Fig. 6, 7, 8 show more results on cross generation setting similar to Fig. 5 of the main paper.

2.2. Video Results

Please find more video results at https://soumik-kanad.github.io/diff2lip. We provide multiple pages of video result comparisons with other methods as well as the original video. The website also contains an interactive demo to see the intermediate stages of the diffusion process. We also show in the wild results using clips from movies dubbed in foreign languages. This along with the cross generations shows the generalizability of the proposed approach to unseen audios and identities. It also supports our motivation and vision to extend videos to different languages beyond simple dubbing which does not have synchronised lip movements. Our method is able to generate realistic lip-movements aligned with the words corresponding to new audio thus making the viewing experience better.

3. Additional details

We have provided the filelists for the subsets of VoxCeleb2 [1] videos on which we have evaluated for both reconstruction and cross generation settings.

Each line contains the relative path to the audio and the video separated by a space. For the case of reconstruction, both the audio and video paths are the same in a line.

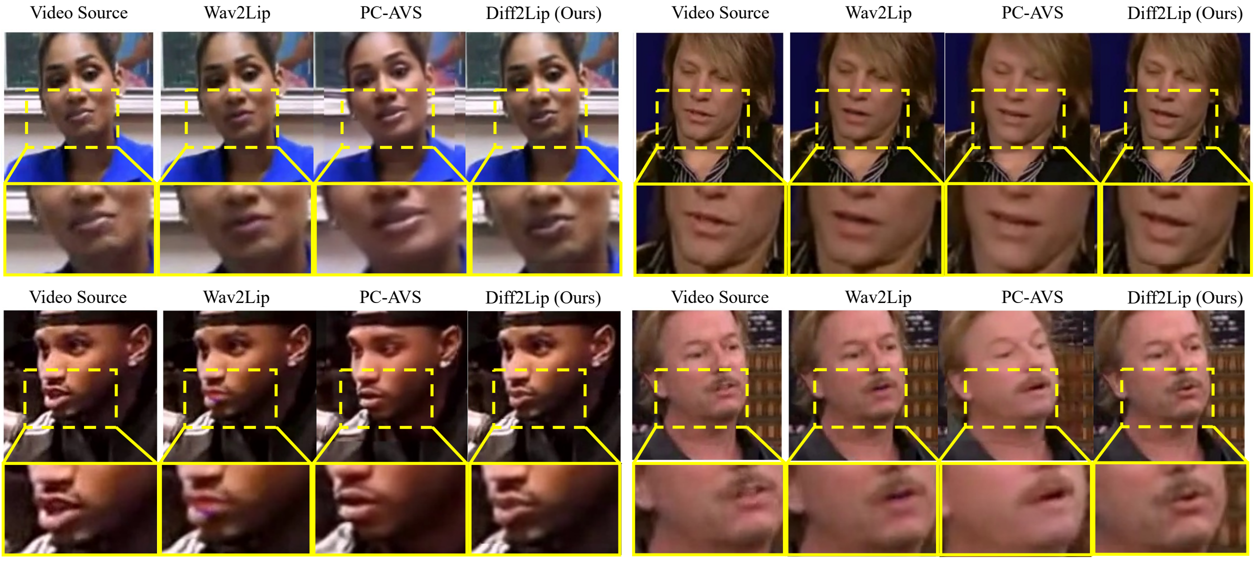

Figure 1. **Qualitative Visual-quality Comparison**.

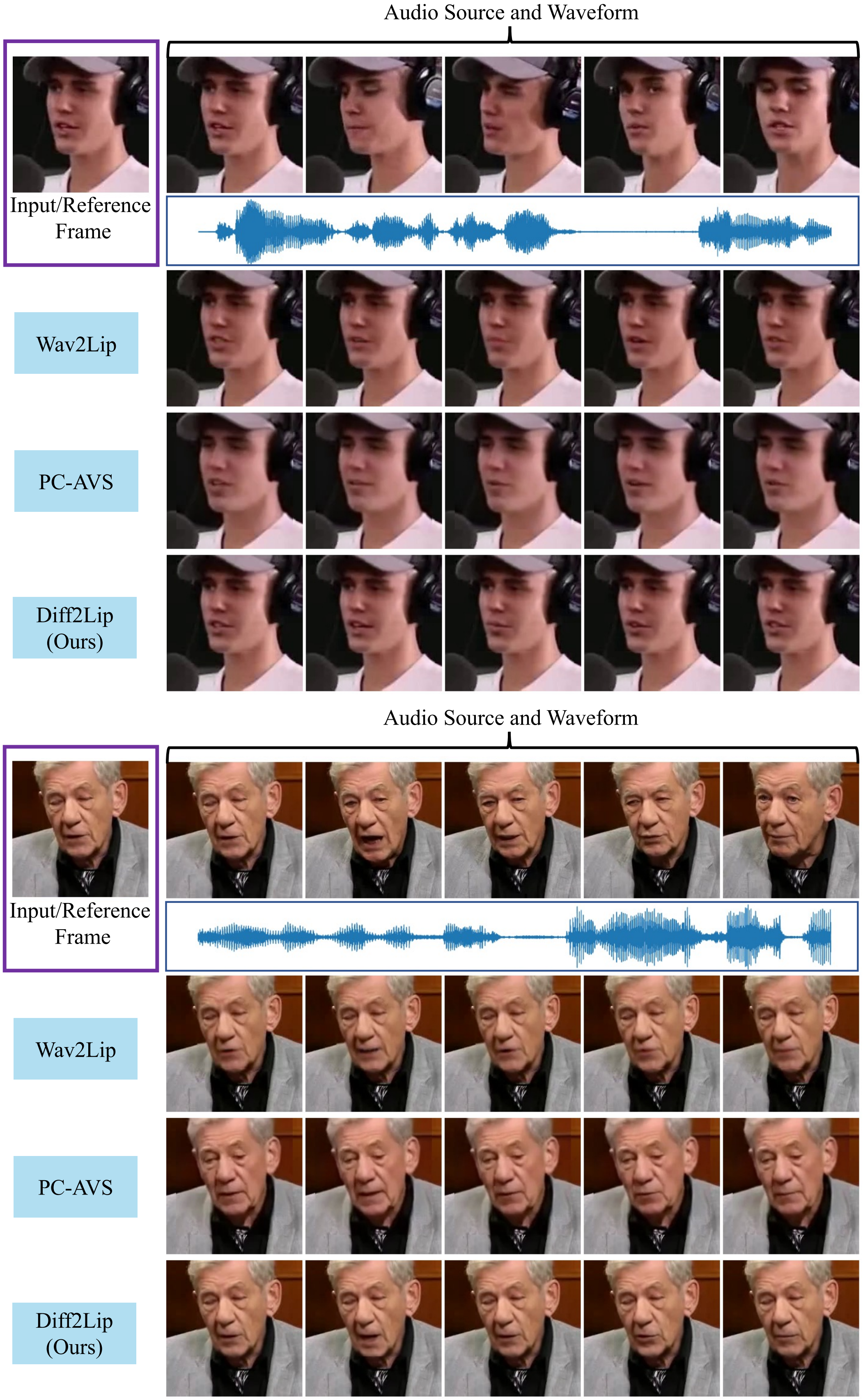

Figure 2. **Qualitative results of Reconstruction on VoxCeleb2 [1].**

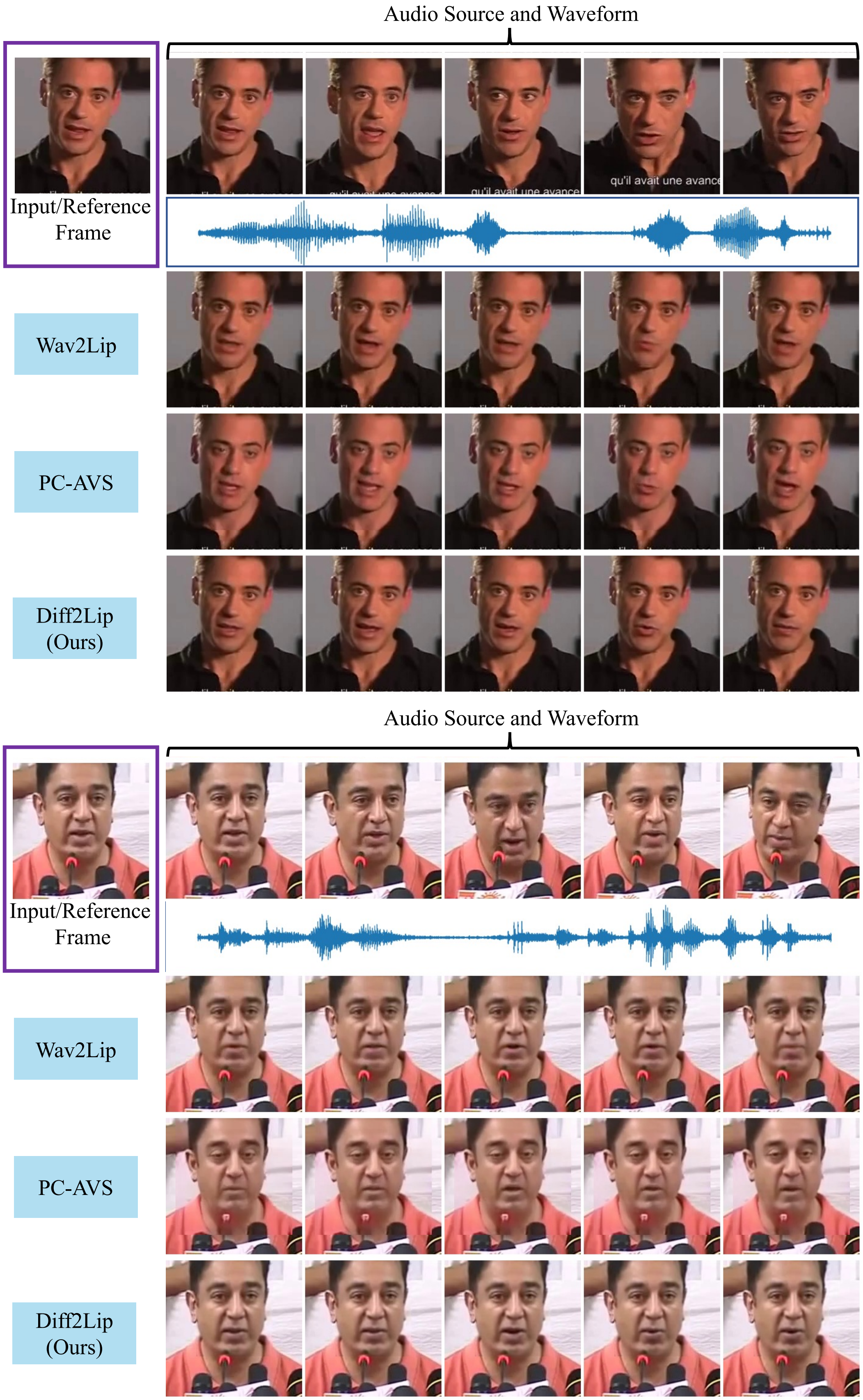

Figure 3. **Qualitative results of Reconstruction on VoxCeleb2 [1].**

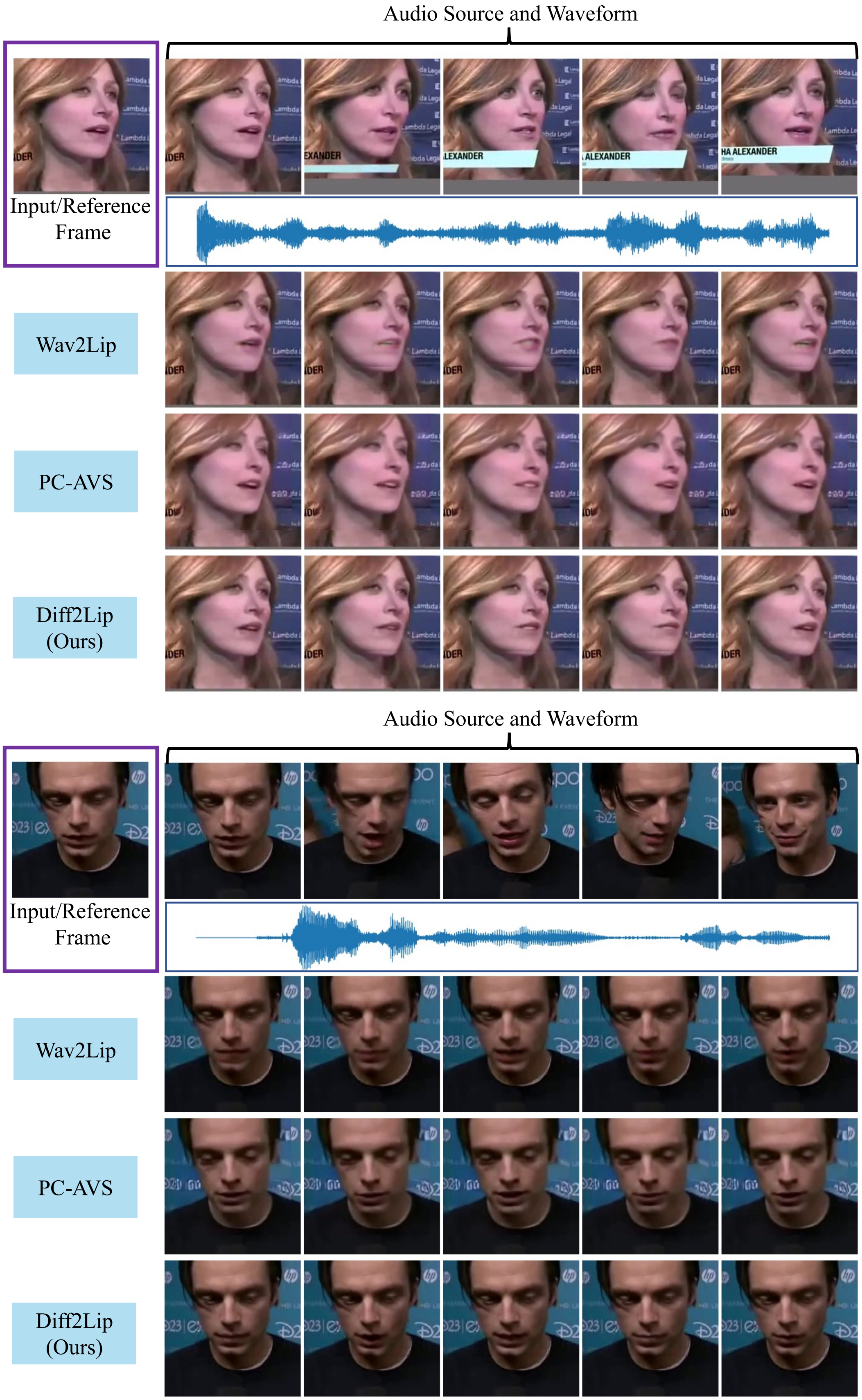

Figure 4. **Qualitative results of Reconstruction on VoxCeleb2 [1].**

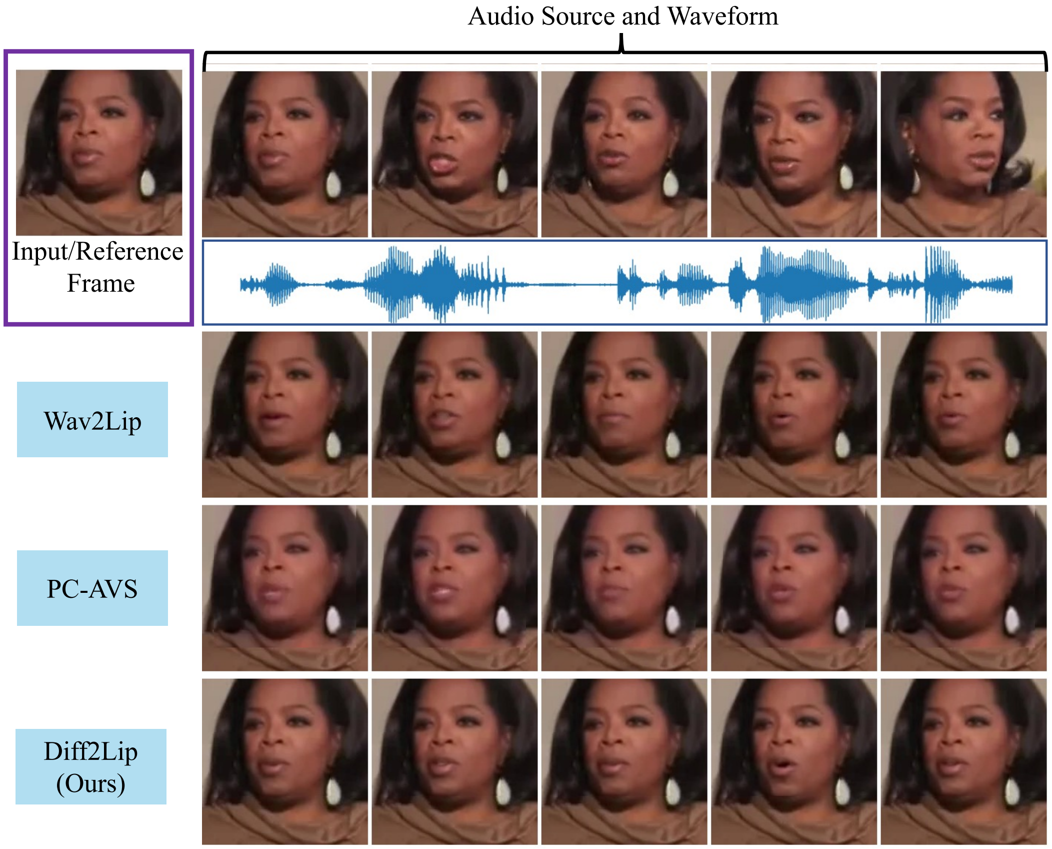

Figure 5. **Qualitative results of Reconstruction on VoxCeleb2 [1].**

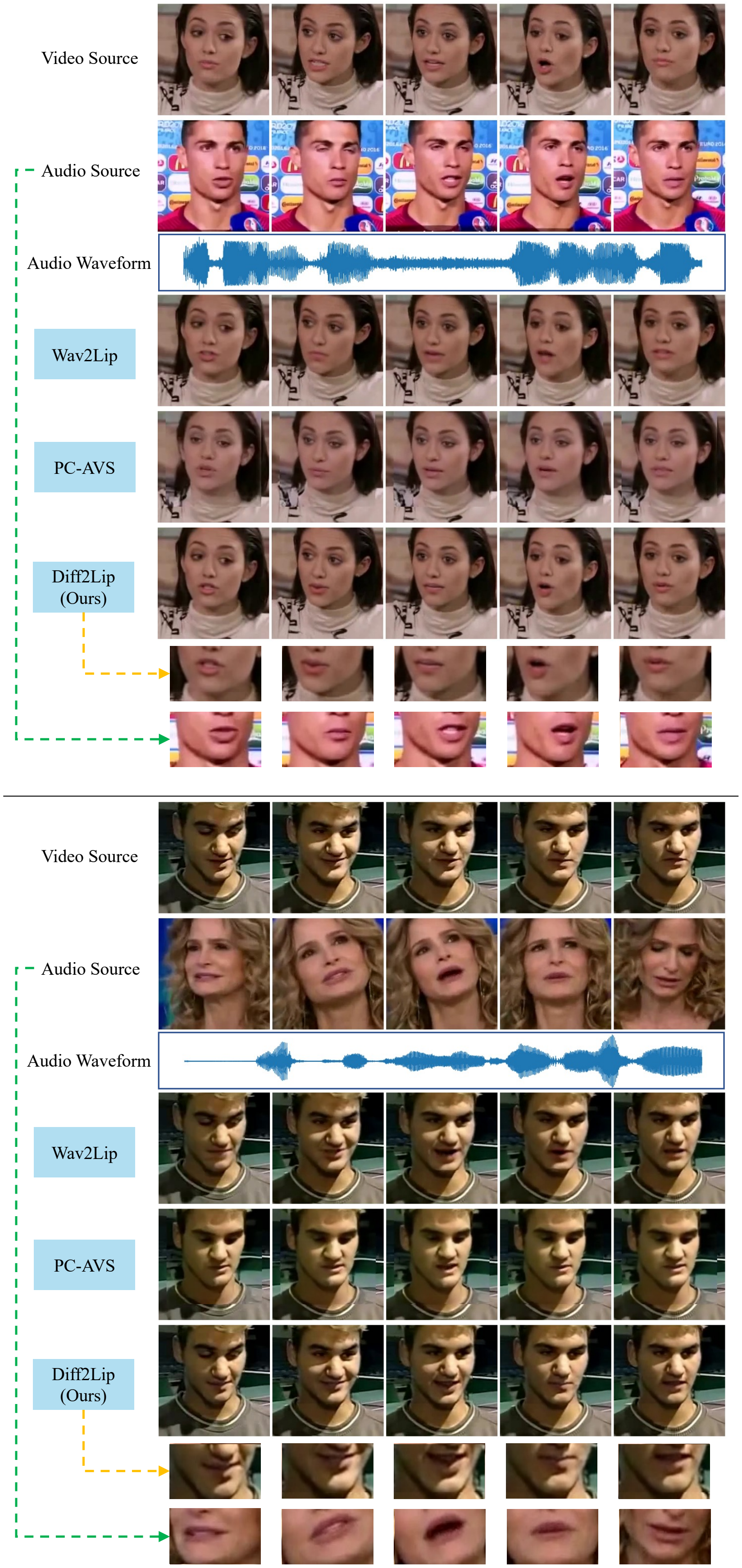

Figure 6. **Qualitative results of Cross generation on VoxCeleb2 [1].**

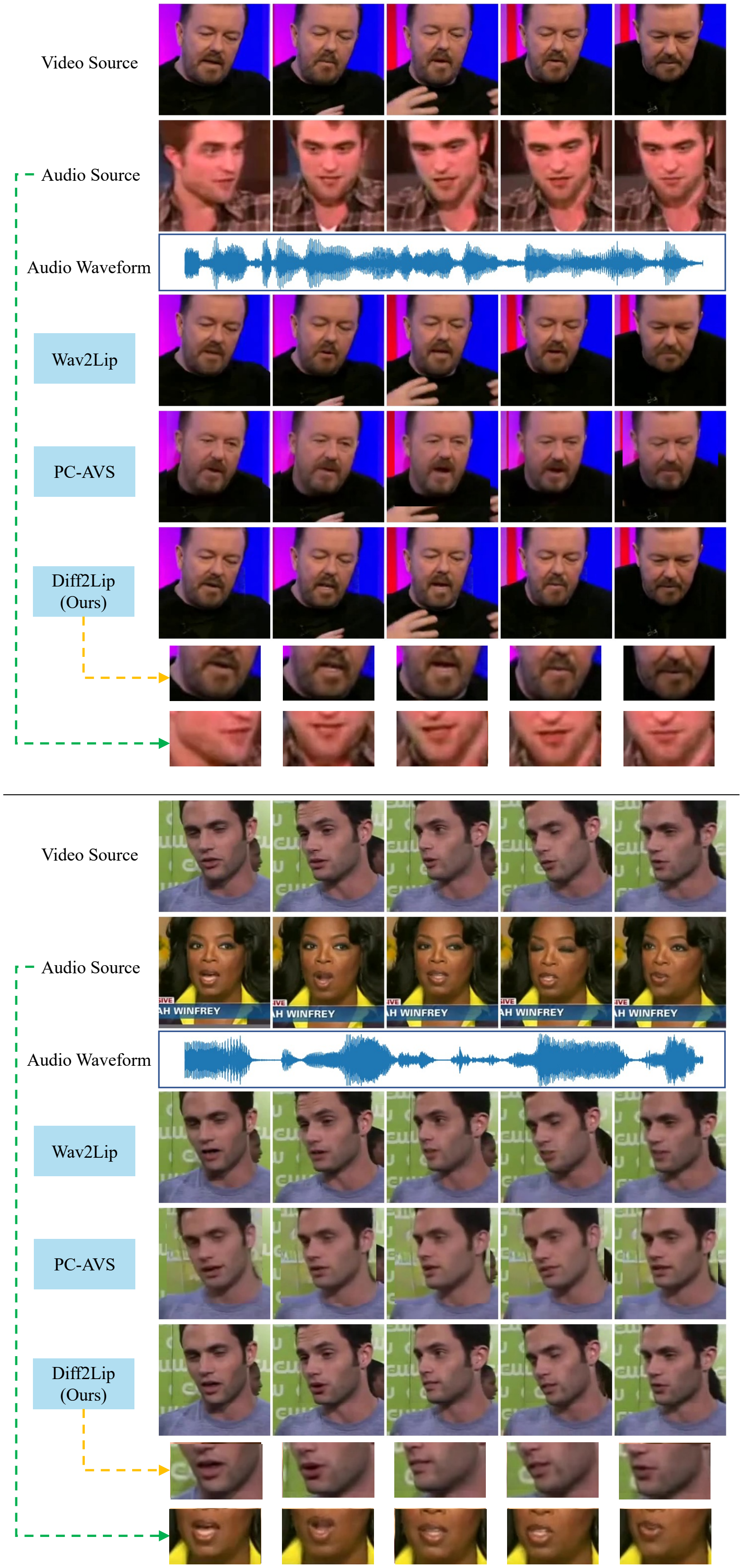

Figure 7. **Qualitative results of Cross generation on VoxCeleb2 [1].**

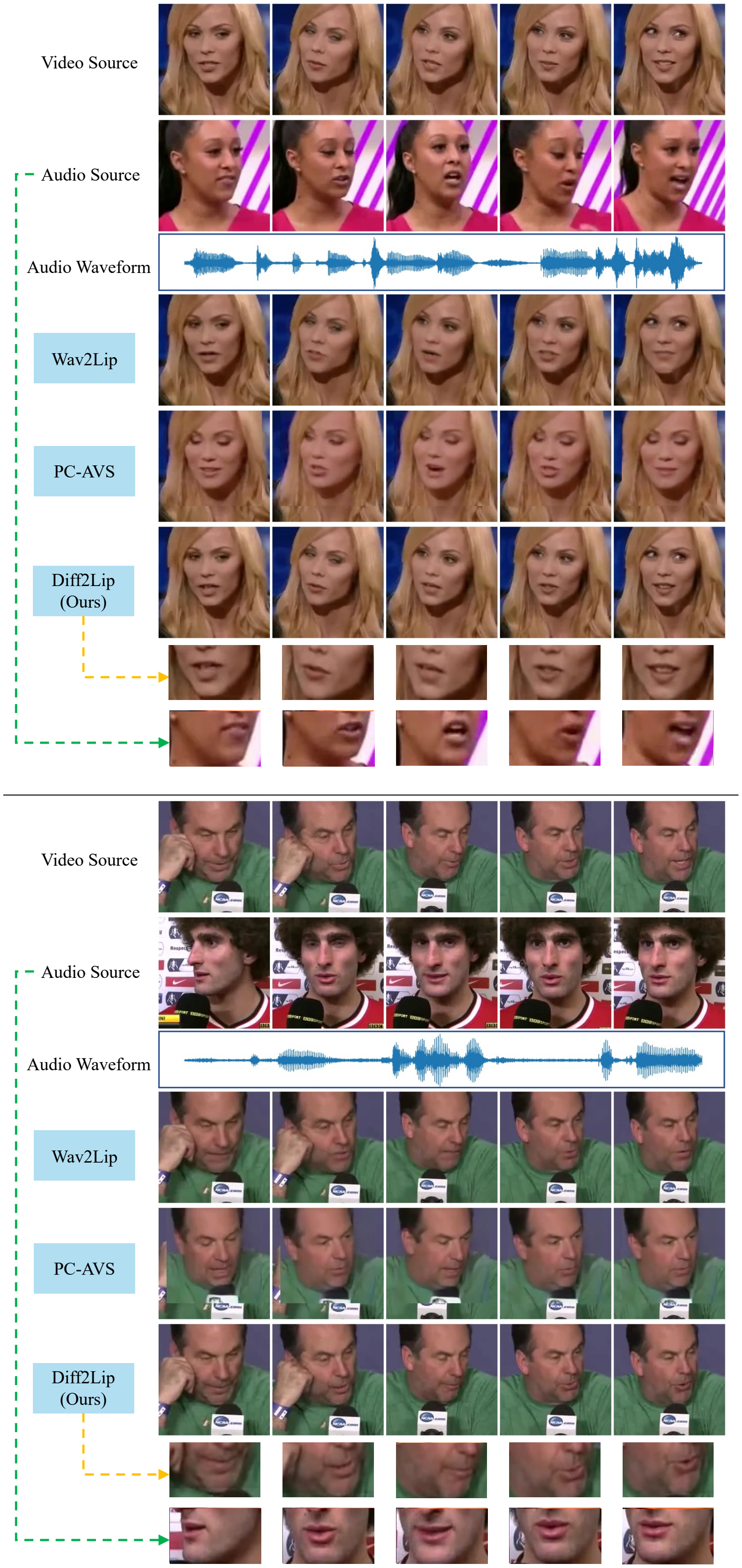

Figure 8. **Qualitative results of Cross generation on VoxCeleb2 [1].**

References

[1] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. Voxceleb2: Deep speaker recognition. *arXiv preprint arXiv:1806.05622*, 2018. 1, 3, 4, 5, 6, 7, 8, 9

[2] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 1

[3] KR Prajwal, Rudrabha Mukhopadhyay, Vinay P Namboodiri, and CV 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*, pages 484–492, 2020. 1

[4] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 1

[5] 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