Supplementary Material for WACV 2026
SynchroRaMa : Lip-Synchronized and Emotion-Aware Talking Face Generation via Multi-Modal Emotion Embedding

Phyo Thet Yee[1] Dimitrios Kollias[2] Sudeepta Mishra[1] Abhinav Dhall[3]
[1]IIT Ropar [2]Queen Mary University of London [3]Monash Universiy
{phyo.22csz0009, sudeepta}@iitrpr.ac.in d.kollias@qmul.ac.uk abhinav.dhall@monash.edu

Figure 1. Given a reference image, audio, and a textual description, Synchrorama can generate talking face videos featuring lip-synchronized, expressive facial expressions and emotional cues while maintaining identity consistency.

This supplementary material provides additional video results on different types of images, such as virtual avatars, real actors, and movie scenes that include other subjects apart from the main character. We also present expressive video results.

Please use Adobe Reader to play the videos. You can find more results at https://novicemm.github.io/synchrorama.

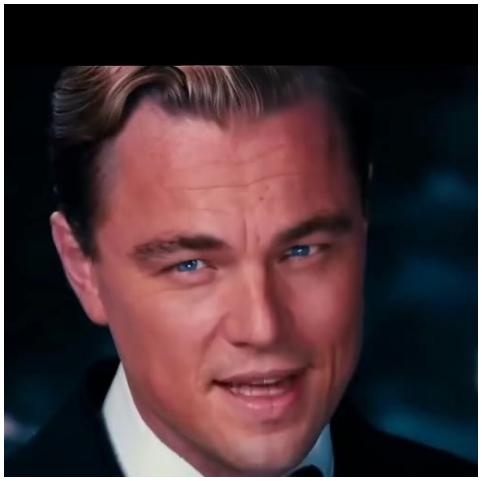

Figure 2. Generated result on a real actor.

Figure 3. Generated result on a real actor.

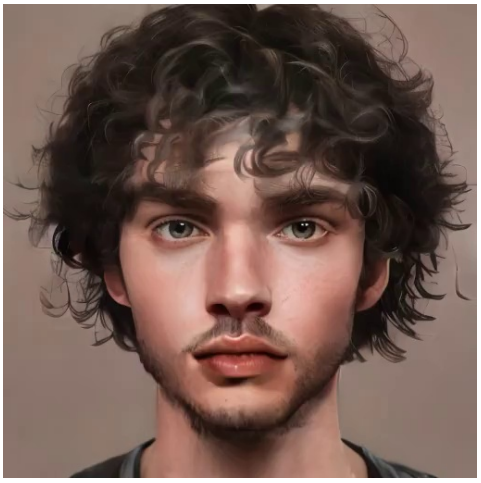

Figure 4. Generated result on a virtual avatar.

Figure 7. Generated result on a movie scene.

Figure 5. Generated result on a virtual avatar.

Figure 8. Generated result with happy expression.

Figure 6. Generated result on a movie scene.

Figure 9. Generated result with fear expression.

Failure Cases: Although our model works well in most cases, it struggles in some situations, such as noisy audio and extreme emotions. We train the model on clean audio, where we remove background noise and music before extracting audio features, Valence Arousal (VA) values, and emotion embeddings. So, the audio and emotion modules mainly learn from clean inputs. When the test audio is noisy (for example, loud background noise or another person speaking), the model receives a noisy conditioning signal. This makes the extracted audio features less reliable and causes small lip-sync errors. For extreme emotions, our training data contains mostly mild expressions, and only a few clips with very strong emotions. Therefore, it struggles to preserve extreme emotions, which require strong facial movements, such as very wide mouth openings, strong brows, and nose wrinkles. In future work, we plan to address these limitations using more robust audio features and adding more training samples with extreme emotions.