Emotionally Enhanced Talking Face Generation
Supplementary Material

Sahil Goyal
IIT, Roorkee, India
sahil_g@ma.iitr.ac.in

Sarthak Bhagat
Carnegie Mellon University
sarthakb@andrew.cmu.edu

Shagun Uppal
Carnegie Mellon University
shagunu@andrew.cmu.edu

Yi Yu
NII, Japan
yiyu@nii.ac.jp

Yifang Yin
A*STAR, Singapore
yin_yifang@i2r.a-star.edu.sg

Rajiv Ratn Shah
IIIT Delhi, India
rajivratn@iiitd.ac.in

1. Related Work

We review the work done in talking face generation and how human emotion is utilized in generating realistic talking face videos separately as follows.

1.1. Talking Face Generation

Several recent works focused on generating talking face videos using deep neural networks. Wu et al. [25] proposed ReenactGAN for talking face generation using the face reenactment technique, which helped transfer the facial landmarks and expressions from a source video of an arbitrary person to the target identity. The landmark boundary encoding was extracted from an arbitrary person's video and mapped to the target person's video via a decoder. Some other works [11, 27] also used facial landmark-based face reenactment techniques for generating video frames. Chen et al. [4] used facial landmarks and a cascade GAN approach to generate desired videos. In this approach, the audio embedding was transferred to facial landmarks, which were then used to generate videos using a regression-based discriminator. Zhang et al. [26] proposed Facial-GAN, which considered explicit face attributes like lip movements and implicit face attributes such as head pose, and eye blink to generate high-quality video frames. Video-based methods that modified only the lip region of the face [14, 18, 24, 16, 21] can generate high-quality talking face videos. They copied the upper half of the face from the input video to generate the target video and hence could not modify the facial expressions and emotions in the upper half of the face. These works did not use human emotion in their models, one of the most critical explicit attributes the model should incorporate to generate more realistic talking face videos.

1.2. *Emotional* Talking Face Generation

Earlier methods [20, 5] tried to infer facial emotions implicitly from audio. However, they have not succeeded at accurately reproducing realistic animation and have struggled to control facial expressions. In contrast, We explicitly feed the desired emotion category as the model input.

Ji et al. [12] proposed an *Emotion Video Portraits* (EVP) algorithm to incorporate the emotion of the audio signal within the target video. Using a Cross-Reconstructed Emotion Disentanglement technique, they decomposed the audio input into a duration-dependent content feature and a duration-independent audio feature. With these two features, emotional facial landmarks were extracted. They introduced the Target-Adaptive Face Synthesis technique that adapted the inferred facial landmarks to the target video. However, they relied on intermediate global landmarks (or edge maps) to generate textures with emotions and on an additional Dynamic Time Warping [2] algorithm to develop their training data to enable cross-reconstructed training. They obtained the latent emotion representation from the input audio using audio emotion disentanglement and then used that disentangled emotion as an explicit modality. Hence, the disentanglement accuracy determined the control of the emotion, making it challenging to have flexible and fully independent control of the emotion. Wang et al. [22] proposed an emotional talking face generation method with explicit emotion control and MEAD dataset (a diverse emotional audio-visual dataset). Similar to our method, they used one-hot representation for emotion. However, they proposed a two-branch architecture, one branch for modifying only the upper half of the face based on emotions and the other for modifying only the lower half of the face using an LSTM [10]-based audio-to-landmarks module. This resulted in inconsistent and conflicting emotions on the face. So unlike the above-discussed methods, our

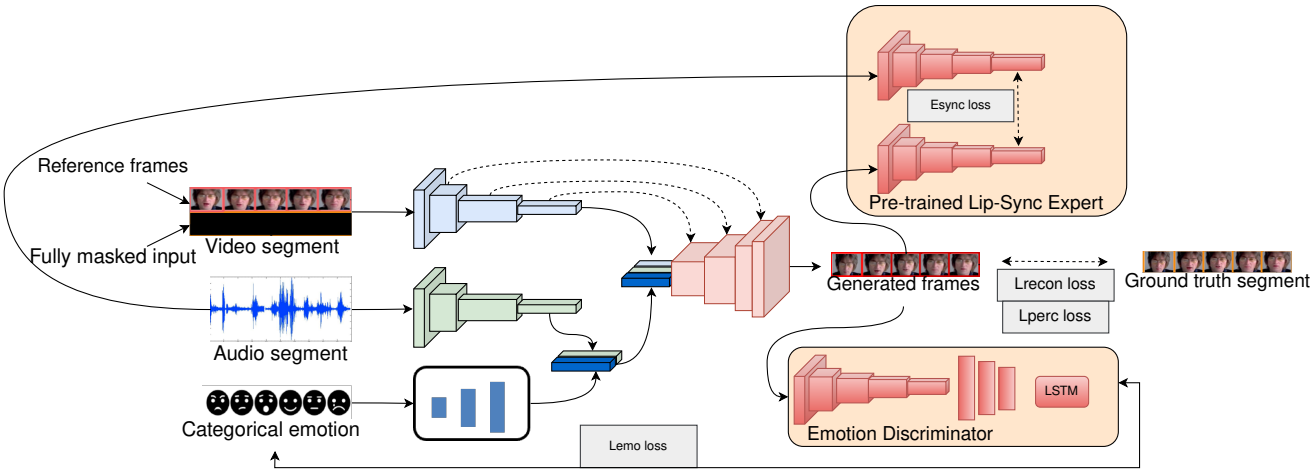

Figure 1: We illustrate a video generation end-to-end network built upon base skeleton architecture. It accepts a continuous set of frames (fully masked) concatenated with reference frames, the Mel spectrogram form of a speech utterance, and a categorical emotion. We concatenate their embeddings in a skip-connection style as shown in this Figure to generate a lip-synced video rendered with the input emotion.

work incorporates emotions into the whole face and uses an audio-independent emotion to generate the talking face videos. Also, EVP [12] and MEAD [22] were involved in training target-specific texture models. Their work is based on single-identity generation. So unlike our model, they perform well only on the subject they are trained on and cannot adapt to arbitrary identities.

Magnusson et al. [13] modified the architecture proposed in [14] to modify emotion using L1 reconstruction and pre-trained emotion objectives. However, their work suffered from several limitations. They did not modify the audio of the source video but retained the original one, which is not the case in most practical applications. In contrast, our model can choose arbitrary audio, ensuring lip synchronization accordingly. Also, their model only modified emotion between specific pairs of emotions (`happiness`, `sadness`, and `neutral`), whereas our model has a broad range of six categorical emotions. Moreover, they trained separate models for each type of emotion transfer. In contrast, our single model can handle all kinds of emotion transfers.

Most models that allow emotion control are image-based models [20, 22, 5, 9, 15] (i.e., which use an identity image as an input along with speech utterance), hence rendering only minor head movements and produce low-quality results. They cannot be used in real-world scenarios. Existing work in emotional talking face generation is limited (especially in the case of video-based models). To the best of our knowledge, this is one of the first studies in which the expression and emotion of a person are considered to generate lip synchronization and talking face generation from video input.

2. Experiments

This section discusses the dataset utilized, methods for concatenating embeddings, implementation details, and our experimental findings.

Figure 2: Augmented frame of an example of CREMA-D [3] dataset. The leftmost is the reference frame, followed by fully masked input, generated frame, and the ground truth frame.

2.1. Dataset

To incorporate the emotions, a dataset with emotion labels is required, and according to our approach, it should fulfill the requirement of a single face in every frame of each clip. Currently, only a few such datasets are publicly available. We use CREMA-D for our purpose. Here are the main attributes of the dataset:

- It contains 7442 clips from 91 actors (48 male and 43 female).
- Actors spoke from a selection of 12 sentences.
- Sentences were presented using one of the six emotions (`happiness`, `sadness`, `fear`, `anger`, `disgust`, `neutral`).

- The image resolution of the clips is 480×360.

We arbitrarily select 5 out of 91 actors and used all of their videos as the test dataset, and the rest of the videos formed the training dataset. We also employ several data augmentation techniques on our input frames, such as random brightness contrast, random Gamma, channel shuffle, RGB shift, and Gaussian noise to generalize our model better. The same augmentations were used in all the input frames to make the frames consistent in visual features like background color, contrast, luminance, brightness, etc. This helped us increase the training data and helped our model generalize over the different background settings. See Figure 2 for an example.

2.2. Concatenating Methods

We try to concatenate the emotion embedding to video and speech embedding using two approaches:

End Concatenation (`END`). We concatenate the emotion encoding at the final step with the video and audio encoding already concatenated. For this, we repeat the emotion $T = 5$ (number of frames per input) along the first dimension. Then after passing through the emotion encoder, we get a latent representation of emotion which is then concatenated with already concatenated audio and face embeddings and is eventually passed through the final output block to get the generated frames of the video.

$$\underset{\substack{\text{Already concatenated}\\ \text{face and audio embeddings}}}{\{N*T,80,96,96\}} + \underset{\text{Emotion embedding}}{\{N*T,1,96,96\}} \equiv \underset{\text{Final embedding}}{\{N*T,81,96,96\}}$$

N, T are batch size and the number of input frames. Note that to concatenate the audio and video embeddings, we process them through face decoder blocks using skip connections (from outputs of layers of different resolutions of face/video encoder blocks).

Sequential Concatenation (`SEQ`). We concatenate the emotion encoding through skip connections similar to the audio encoding. We first concatenate the audio and emotion embedding. The concatenated embedding is processed through face decoder blocks of the generator using skip connections along with face embedding as shown in Figure 1.

2.3. Noise Encoder

We introduce a noise encoder in the initial part of our model, along with a face, audio, and emotion encoder. A noise vector is drawn from the standard Gaussian distribution for each video frame. We process this sequence of noise vectors through a single layer of an LSTM [10] encoder to get noise embedding which is concatenated with the face embeddings. The motive for introducing this module of temporal noise is to account for randomnesses, such as head movements and eye blinking, independent of the input data. We do not incorporate a noise encoder in any of our experimental settings.

2.4. Pre-training the Base Model (`PRE`)

LRS2 [1] is relatively larger than CREMA-D [3] and has more complex head poses, but it cannot be used for our modified model because it does not have categorical emotion labels. Hence, we try to pre-train the base model (that does not require emotion labels) on the LRS2 dataset and then use the face encoder block from the pre-trained model in two ways (as the architecture of the face encoder is the same in both the base model and the modified model):

- Keeping the weights of the face encoder fixed while training the modified model.
- Using pre-trained weights of face encoder as initialization for training the modified model.

We also modify the base model to generate the whole face instead of only the lip region and then pre-train it.

2.5. Implementation Details

Adam optimizer [8] is used for training all the networks with β_1 and β_2 as 0.5 and 0.999 respectively. The learning rate for updating the emotion discriminator and generator is $1e^{-6}$ and $1e^{-4}$, respectively. The full objective function of training the generator is

$$\begin{aligned} L_{gen} = \alpha E_{sync} + \beta L_{perc} + \gamma L_{emo} \\ +(1-\alpha-\beta-\gamma)L_{recon} \end{aligned} \tag{1}$$

where, α, β, γ are the weights for the respective loss components. Constant α is set to 0 initially and later updated to 0.03 when the sync-loss on validation data becomes less than a predefined value. β, γ are 0.01 and 0.001 respectively. Images are normalized between the 0 and 1 value range.

By increasing the weight assigned to the emotion loss term, the model can more effectively incorporate emotions into its predictions at an earlier stage of the training process, but it comes at the cost of a slight reduction in reconstruction quality.

3. Inference Details

The Number of frames in the input to the model is a hyperparameter. While training, we set it to 5, and our model generates the 5 new frames corresponding to those input frames, whereas our model inference allows any number of frames to be generated depending upon the duration of audio and video input. More information is as follows:

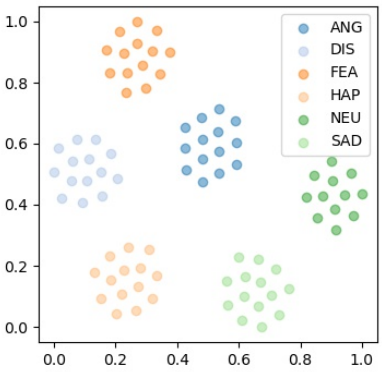

Figure 3: Visualization of the projected emotion embeddings. Each color represents a specific emotion.

Input audio duration = Input video duration. The number of frames generated is the same as the number of frames in the input video.

Input audio duration < Input video duration. During inference, we trim the video, limiting it to the duration of input audio, and the model generates the number of frames equal to that in the trimmed video.

Input audio duration > Input video duration. We repeat the last frame of the input video to extend the input video to the duration of the input audio, and the model generates the number of frames equal to that in the extended video.

4. Ablation Study

We study the efficacy of our different experimental settings in this section.

4.1. Emotion Encoder

We visualize the embeddings learned by our emotion encoder. We use *t-SNE* [19] algorithm to project the learned encodings to a 2-dimensional space as shown in Figure 3. We arbitrarily select ActorID 1011 from the test split of the CREMA-D [3] dataset. We utilize all the videos of that actor for our purpose. We average the embeddings across the frames for each video. Each data point in Figure 3 represents averaged embeddings of a video of ActorID 1011. Clusters formed for different emotions in Figure 3 show that our emotion encoder learns useful representations for the emotion.

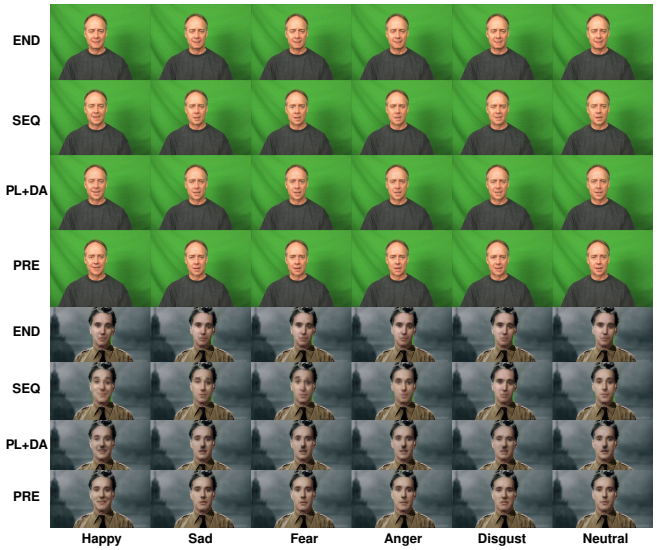

Figure 4: We generate videos for all six emotions and concatenated the specific frames from each. Each row represents an experiment mentioned in Section 4, and each column represents a particular emotion in all the experiments.

4.2. `END` Concatenation

See Section 2.2 for details of the `END` concatenation. We do not employ perceptual loss and data augmentation in this experiment. Although the sync quality is good, the visual and emotional rendering are unsatisfactory. See rows labeled `END` in Figure 4. Moreover, some undesirable green background is present in the frames of the second example because all the training examples have a green screen in their background, so the model cannot generalize completely on other videos. Some arbitrary black dot artifacts are also visible on the generated frames. A possible explanation for the same could be that the one hot emotion vector is sparse. We repeat it for every frame and process this sparse vector formed through an emotion encoder to generate a large tensor, concatenating it to already concatenated audio and video embeddings to generate the required video. So the presence of large-sized sparse matrices in this approach results in black dot artifacts on the frames.

4.3. `SEQ` Concatenation

See Section 2.2 for details of the `SEQ`uential concatenation. This method improves the visual quality and emotional rendering to a large extent. Here, we do not employ perceptual loss or data augmentation. See rows labeled `SEQ` in Figure 4. Emotion is rendered to some extent in the frames. The model still doesn't generalize, as a green background can be seen in the frames. However, those black dot artifacts disappear using the method `SEQ` because this approach reduces the size of the sparse matrices involved.

Figure 5: An example comparing generated frames using a **cartoon subject** sampled from the internet. We chose this subject to evaluate the ability of different approaches to generalize to arbitrary identities. Every fifth frame of the generated video is shown in each row. Wang et al. [23] (second row) completely failed to generate any meaningful video and instead generated frames full of artifacts. Eskimez et al. [9] was unsuccessful in detecting the relevant face from the video in the initial step and could thus not generate an emotional talking face video. Furthermore, Magnusson et al. [13] cannot generate a video for *anger* emotion. In contrast, our approach PL + DA successfully detected the relevant face to generate the realistic frames and effectively conveyed the *anger* emotion on the subject's face.

Figure 6: An example comparing generated frames using a subject from the test dataset of CREMA-D [3]. Every fifth frame of the generated video is shown in each row. The top row corresponds to the ground truth video. Our baseline Prajwal et al. [14] (third row) generated realistic frames but cannot incorporate emotions. Wang et al. [23] (second row) again failed to preserve the subject's identity, resulting in non-human-like faces. Eskimez et al. [9] (fourth row) could not effectively synthesize the *disgust* emotion. Magnusson et al. [13] involves only three emotions (*happiness*, *sadness*, *neutral*). It cannot generate video for *disgust* emotion. In contrast, our approach PRE was able to generate realistic frames that accurately depicted the *disgust* emotion on the subject's face.

This concatenation method is our preferred approach, and we conduct the following experiments using it.

Efficacy of including Perceptual Loss and Data Augmentation (PL+DA). This approach is: SEQ + Perceptual loss + Data Augmentation. See rows labelled PL+DA in Figure

4. We observe the most satisfactory results under these experimental settings. Data augmentation solves the issue of a green background, aiding the model generalizing on videos other than training examples. Also, penalizing the model with perceptual loss improves visual quality and emotion rendering.

Efficacy of Pre-training the Base Model (PRE). This approach is basically: (PL+DA) + pre-training. See Section 2.4 for details of this experiment. See rows labeled PRE in Figure 4. The results show a slight improvement in the frames' visual quality, but a degradation in the temporal continuity of the generated frames is observed. Emotion rendering is comparable to PL+DA.

5. Additional Qualitative Results

See Figure 5 and 6 for additional qualitative results.

6. Quantitative Evaluation Methods

This section discusses the methods used for quantitative evaluation in detail.

6.1. Emotion Incorporation

We exploit an emotion classifier to evaluate the generated emotional talking face videos. We utilize the same architecture as the emotion discriminator in our main pipeline. We trained the classifier for the train split of the CREMA-D [3] dataset. We obtain an accuracy of more than 90% on the test set of the CREMA-D dataset, indicating that our video-based emotion classification model can fairly evaluate the emotion incorporation ability of our model. The higher the emotion classification accuracy (*EmoAcc*) of the video-based emotion classifier, the better the emotion incorporation ability of the model. As we are using arbitrary emotions to generate our videos, those arbitrary emotions can be exploited as ground truth labels for the classifier to evaluate our model.

6.2. Sync Quality

We use the metrics *LSE-C* and *LSE-D*, proposed in [14] to evaluate the sync quality. The lower the *LSE-D*, the higher the sync quality. The higher the *LSE-C*, the higher the sync quality. We use the videos from the CREMA-D [3] dataset, but the audio inputs are randomly sampled from the internet in English and Hindi. All our experiments (END, SEQ, PL + DA, PRE) have a sync quality comparable to our baseline (Wav2Lip [14]) and better than other related works, which means that adding emotion to the base model does not compromise the sync quality.

6.2.1 Calculating *LSE-C* and *LSE-D*

Pre-trained SyncNet released by [7] is utilized to measure the lip-sync error between the generated frames and the randomly chosen speech segment. This SyncNet differs from the *expert lip-sync discriminator* we have used in training. Its architecture is based on Siamese networks [6] and is trained on a public dataset (derived from the BBC videos) using contrastive loss. The pre-trained model is available publicly[1].

A sliding-window technique is utilized to calculate the *LSE-C* and *LSE-D* metrics. For each video clip, multiple samples are extracted because there may be samples in which no one is speaking at that particular time. The Euclidean distance between one 5-frame video feature and all the audio features in the ± 1 second range is calculated for each sample. Then those distances are averaged across all the samples. Out of all those average distances, the minimum one is defined as the Lip Sync Error - Distance (*LSE-D*) because the correct offset is when the distance is minimum. The difference between the median and minimum (*LSE-D*) of the average distances calculated above is defined as the Lip Sync Error - Confidence (*LSE-C*).

6.3. Visual Quality

We use Fréchet Inception Distance (*FID*) for evaluating the visual quality. Feature representations of the two sets of images are encoded using a pre-trained Inception network [17], and then Fréchet distance is calculated between the Gaussian distributions fitted to those representations. The *FID* scores for all the approaches involving emotions are averaged over the six emotion categories. The *FID* for our approach (involving emotion) is expected to be higher than the approaches not involving emotions [14, 23] because emotion incorporation, along with lip synchronization, requires more image manipulation than the ones involving only lip synchronization. The methods not incorporating emotions generate only the lower half region of the face, i.e., the lip region, whereas, for emotion incorporation, we generate the entire face. However, our visual quality improved significantly due to the addition of perceptual loss in PL+DA and PRE settings. The significant difference between PL+DA and PRE settings is additional knowledge gained by PRE through pre-training. The PRE approach outperforms all other methods in *FID*.

7. Web Interface

Our proposed framework includes a user-friendly web interface that allows users to generate talking faces with emotions using the model with PL+DA settings. Currently, the model uses an NVIDIA TITAN Xp GPU for inference.

[1] https://github.com/joonson/syncnet_python

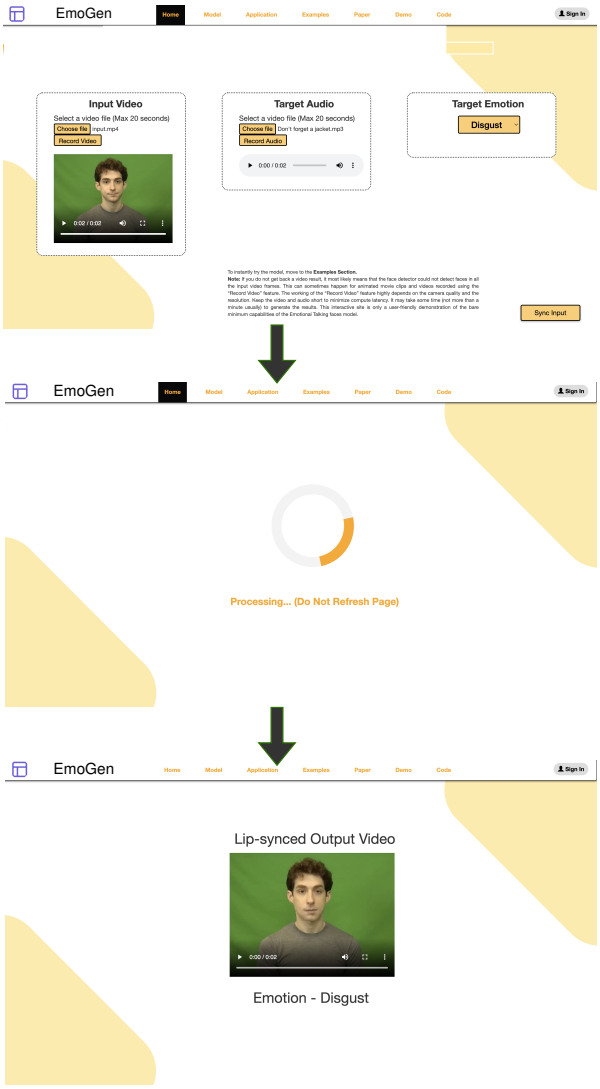

Figure 7: Working of the demo website.

FastAPI (Python Framework) is used for the backend development of the interface, which handles all the API requests. HTML, CSS, and Javascript are used for front-end development. All the clients' requests are sent to the backend via Javascript using a fetch call. Request details are sent in JSON format. The website is hosted on HTTPS to address security issues. The website is super-easy to use, as illustrated in Figure 7. Following are some basic steps to use the demo website:

- Before using the interface, read the instructions on the home page.
- Choose an arbitrary video, audio, and emotion as inputs. You can also use the recording feature for video and audio inputs. Then press the "Sync Input" button (located at the bottom right of the home page).

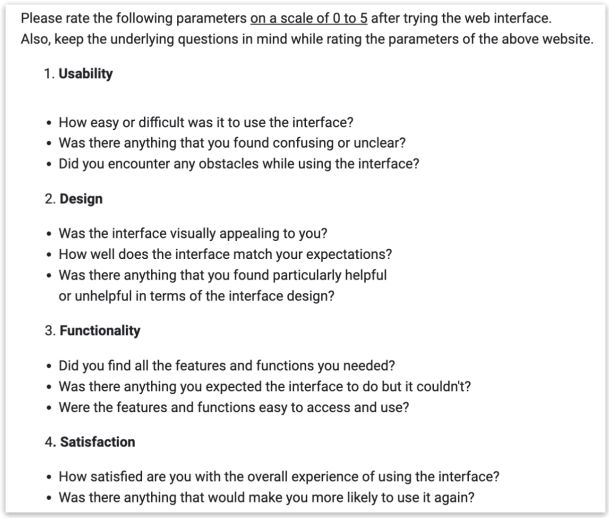

Figure 8: Details of the user study.

- After a 20 to 30 seconds wait, the emotionally enhanced and lip-synced talking face video will be ready.

A video illustration is also provided in the accompanying supplementary materials

8. User Study

We conduct a user study through subjective evaluation to understand the user experience on our web interface. We survey a diverse group of 25 users about their experience navigating and using the website. We ask them to rate the ease of usability, design, functionality, and overall experience on a scale of 0 to 5. The higher the rating, the better the web interface. Additionally, we ask them for specific suggestions. We also provide them with a small illustration video explaining some basic steps to use the website. See Figure 8 for more details. The user study results provided valuable insights into the strengths and weaknesses of our web interface. The feedback from the participants will enable us to improve the website, particularly its design.

9. Image vs. Video as an Input

Using an image as an input instead of a video will obviously render significantly fewer head movements in the generated video, as video-based models can also inherit head movements from the source input video. Video-based models can generate videos with much better temporal coherence, meaning the movements of the mouth and other facial features are more consistent over time. They can incorporate more subtle variations in facial expression and movement, which makes the videos appear more lifelike and convincing. Moreover, in real-world applications like dubbing

movies, TV shows, etc., We cannot use image-based models because to generate the dubbed video, we will definitely require a source video input.

10. Emotion as an explicit modality

Explicit control over the emotion in talking face video generation can be valuable in various real-world applications. For example, In a video advertisement, the audio may have a neutral tone, while the visuals may need to convey a happy or excited emotion. While extracting emotions from the audio can be helpful, the emotional content of the audio may not be clear or easily separable from other audio signals. Regarding the robustness and consistency of extraction from audio, the emotional content of speech can be affected by various factors such as tone, pitch, and volume, which can be challenging to separate from the speech signal. Also, the existing models may not accurately capture the emotion from audio, leading to inaccuracies in the extracted emotional expression.

11. Limitations and Future Work

Our approach, however, is limited by the availability of datasets with categorical emotion labels that are long enough and have exactly one face in each frame. Our current approach does not allow the use of datasets with multiple faces in a single frame, and the short datasets do not allow the model to generalize effectively. CREMA-D [3] contains relatively simple videos (with only a straight head pose). We can find or collect a better dataset for future work. It should be long enough to make the model generalize better and have videos with different head poses. One such potential dataset is MEAD [22].

Various further improvements can be included in future work. Some better masking methods can be explored to mask the ground truth frames (such as using a convex hull). Different ways to enforce the input emotion on the final audio can be examined, such as using an additional loss function. For evaluating the emotion rendering of the model, deepfake detectors that detect deepfakes based on inconsistency in emotions can be used. Also, some more relevant metric than *FID* score is required to access the visual quality in the case of emotion incorporation because emotion rendering leads to more significant changes in the face than lip synchronization.

12. Ethical Use

Synthetic video generation has many potential applications, including entertainment, education, and marketing. However, their use also raises ethical concerns that must be carefully considered. Talking face generation, videos may spread misinformation or propaganda or impersonate individuals for fraudulent or malicious purposes. It can lead to reputation damage and emotional distress. As these videos become more sophisticated and difficult to detect, it becomes increasingly challenging to distinguish real from fake content. This undermines the integrity of the media. Given these risks, it is essential to consider how synthetic media can be regulated or controlled to minimize their negative consequences. One possibility is developing high-quality algorithms or tools that detect and flag synthetic content. Another approach is establishing legal frameworks or guidelines that outline the acceptable uses of talking face generation videos and penalties for misuse.

Finally, it is crucial to recognize that the creators and users of talking face generation videos are responsible for ensuring that they are used ethically, which includes considering the potential impacts of their work on others and taking steps to minimize any negative consequences. It also involves being transparent about synthetic media and clearly labeling content as manipulated when appropriate.

References

[1] Triantafyllos Afouras, Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Zisserman. Deep audio-visual speech recognition. *IEEE transactions on pattern analysis and machine intelligence*, 2018. 3

[2] Donald J Berndt and James Clifford. Using dynamic time warping to find patterns in time series. In *KDD workshop*, volume 10, pages 359–370, 1994. 1

[3] Houwei Cao, David G Cooper, Michael K Keutmann, Ruben C Gur, Ani Nenkova, and Ragini Verma. Crema-d: Crowd-sourced emotional multimodal actors dataset. *IEEE transactions on affective computing*, 5(4):377–390, 2014. 2, 3, 4, 5, 6, 8

[4] Lele Chen, Ross K Maddox, Zhiyao Duan, and Chenliang Xu. Hierarchical cross-modal talking face generation with dynamic pixel-wise loss. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7832–7841, 2019. 1

[5] Lele Chen, Guofeng Cui, Celong Liu, Zhong Li, Ziyi Kou, Yi Xu, and Chenliang Xu. Talking-head generation with rhythmic head motion, 2020. URL https://arxiv.org/abs/2007.08547. 1, 2

[6] S. Chopra, R. Hadsell, and Y. LeCun. Learning a similarity metric discriminatively, with application to face verification. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 539–546 vol. 1, 2005. 6

[7] Joon Son Chung and Andrew Zisserman. Out of time: automated lip sync in the wild. In *Asian conference on computer vision*, pages 251–263, 2016. 6

[8] John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, 12(7), 2011. 3

[9] Sefik Emre Eskimez, You Zhang, and Zhiyao Duan. Speech driven talking face generation from a single image and an emotion condition. *IEEE Transactions on Multimedia*, pages 1–1, 2021. doi: 10.1109/TMM.2021.3099900. 2, 5

[10] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 1, 3

[11] Po-Hsiang Huang, Fu-En Yang, and Yu-Chiang Frank Wang. Learning identity-invariant motion representations for cross-id face reenactment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 1

[12] Xinya Ji, Hang Zhou, Kaisiyuan Wang, Wayne Wu, Chen Change Loy, Xun Cao, and Feng Xu. Audio-driven emotional video portraits. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14080–14089, June 2021. 1, 2

[13] Ian Magnusson, Aruna Sankaranarayanan, and Andrew Lippman. Invertable frowns: Video-to-video facial emotion translation. In *Proceedings of the 1st Workshop on Synthetic Multimedia-Audiovisual Deepfake Generation and Detection*, pages 25–33, 2021. 2, 5

[14] 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, 2, 5, 6

[15] Sanjana Sinha, Sandika Biswas, Ravindra Yadav, and Brojeshwar Bhowmick. Emotion-controllable generalized talking face generation. *arXiv preprint arXiv:2205.01155*, 2022. 2

[16] Linsen Song, Wayne Wu, Chen Qian, Ran He, and Chen Change Loy. Everybody's talkin': Let me talk as you want. *IEEE Transactions on Information Forensics and Security*, 17:585–598, 2022. 1

[17] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015. 6

[18] Justus Thies, Mohamed Elgharib, Ayush Tewari, Christian Theobalt, and Matthias Nießner. Neural voice puppetry: Audio-driven facial reenactment. In *European conference on computer vision*, pages 716–731. Springer, 2020. 1

[19] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9 (11), 2008. 4

[20] Konstantinos Vougioukas, Stavros Petridis, and Maja Pantic. Realistic speech-driven facial animation with gans. *CoRR*, abs/1906.06337, 2019. URL http://arxiv.org/abs/1906.06337. 1, 2

[21] Ganglai Wang, Peng Zhang, Lei Xie, Wei Huang, and Yufei Zha. Attention-based lip audio-visual synthesis for talking face generation in the wild. *arXiv preprint arXiv:2203.03984*, 2022. 1

[22] Kaisiyuan Wang, Qianyi Wu, Linsen Song, Zhuoqian Yang, Wayne Wu, Chen Qian, Ran He, Yu Qiao, and Chen Change Loy. Mead: A large-scale audio-visual dataset for emotional talking-face generation. In *European Conference on Computer Vision*, pages 700–717, 2020. 1, 2, 8

[23] Suzhen Wang, Lincheng Li, Yu Ding, Changjie Fan, and Xin Yu. Audio2head: Audio-driven one-shot talking-head generation with natural head motion. *arXiv preprint arXiv:2107.09293*, 2021. 5, 6

[24] Xin Wen, Miao Wang, Christian Richardt, Ze-Yin Chen, and Shi-Min Hu. Photorealistic audio-driven video portraits. *IEEE Transactions on Visualization and Computer Graphics*, 26(12):3457–3466, 2020. 1

[25] Wayne Wu, Yunxuan Zhang, Cheng Li, Chen Qian, and Chen Change Loy. Reenactgan: Learning to reenact faces via boundary transfer. In *Proceedings of the European conference on computer vision (ECCV)*, pages 603–619, 2018. 1

[26] Chenxu Zhang, Yifan Zhao, Yifei Huang, Ming Zeng, Saifeng Ni, Madhukar Budagavi, and Xiaohu Guo. Facial: Synthesizing dynamic talking face with implicit attribute learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3867–3876, October 2021. 1

[27] Jiangning Zhang, Xianfang Zeng, Mengmeng Wang, Yusu Pan, Liang Liu, Yong Liu, Yu Ding, and Changjie Fan. Freenet: Multi-identity face reenactment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 1