Efficient Emotional Adaptation for Audio-Driven Talking-Head Generation (Supplementary)

Yuan Gan[1,2], Zongxin Yang[1,2], Xihang Yue[1,2], Lingyun Sun[2], Yi Yang[1,2*]
[1]ReLER, CCAI, Zhejiang University, China
[2]College of Computer Science and Technology, Zhejiang University, China

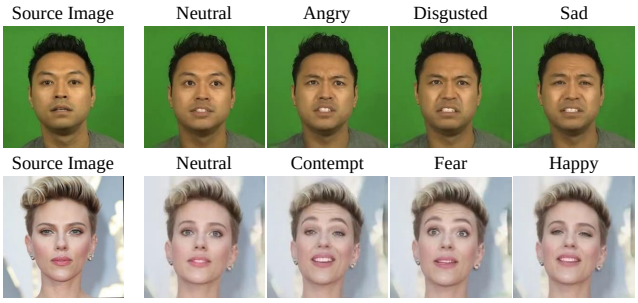


Figure 1. Additional emotional expressions generated by EAT. EAT produces realistic and diverse facial expressions with corresponding emotional guidance. Please zoom in for a better view. Source images are from CREMA-D[3] and MakeItTalk[23].

A. The Networks Details

We provide additional details of our network architecture and training procedure. It should be noted that the Keypoint Detector (D_k) and RePos-Net networks are primarily derived from OSFV [19]. For more information, interested readers may refer to OSFV [19].

Audio-to-Expression Transformer. We use the Audio-to-Expression Transformer (A2ET) to transfer the audio to 3D latent expression deformation sequences. The A2ET consists of an encoder and a decoder, both with 6 transformer layers and 8 heads. The feed-forward layer has a dimension of 1024. Each token is a 128-dim vector. The expression deformation vector (E_i) is predicted by the feature of the central frame i. However, directly optimizing the 3D expression motions leads to convergence problems in network training. To address this issue and bridge the gap between the 3D expression deformation and the audio features, we use principal component analysis (PCA) to reduce the dimensionality of E_i from 45 to 32. Specifically, we calculate the matrix of principal eigenvalues U and mean vector M from the training set. Then the expression deformation vector is obtained by projecting the predicted PCA using the following equation:

$$E_i = PE_i * U^T + M, \tag{1}$$

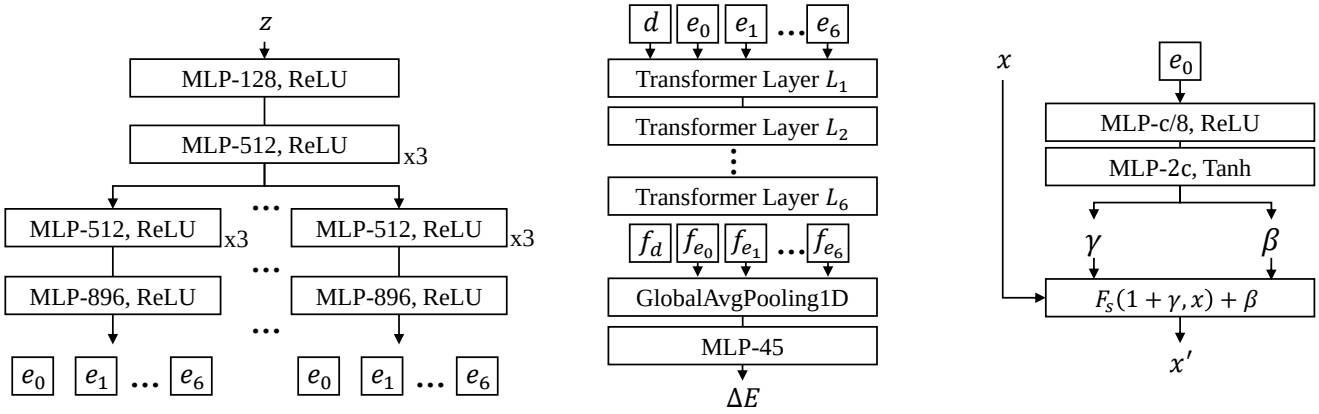
where PE_i is the predicted PCA and E_i is the expression deformation, which is used to modify the neutral 3D keypoints to generate the expressive face. As the number of keypoints is 15, the shape of E_i is (15, 3).

Emotion Mapper. We propose an emotion mapper that produces emotional tokens to guide the generation of emotional expressions. As shown in Fig. 2(a), the emotion mapper M consists of several shared and unshared multi-layer perceptrons (MLP) layers. It takes a 16-dim latent code z as input and outputs seven emotional tokens e_0, e_1, $\cdots$, e_6. The first token e_0 serves as the emotional guidance for the emotional adaptation module (EAM), which modifies the features of the audio-to-expression transformer (A2ET). The remaining six tokens e_1, $\cdots$, e_6 are fed to the corresponding transformer layer of A2ET as deep emotional prompts. The Emotional Deformation Network (EDN) then uses all these tokens and the latent source representation to generate the emotional deformation ΔE.

Emotional Deformation Network. The Emotional Deformation Network (EDN) learns the emotional deformation ΔE using the same architecture as the A2ET encoder, which has six transformer layers. Fig. 2(b) shows the input and output of EDN. It takes the latent source representation d and the emotional guidance tokens $e_0, e_1, \cdots, e_6$ as input, and extracts their features f_d, f_{e_0}, $\cdots$, f_{e_6}. Then it applies global average pooling to the emotion-related features f_{e_0}, $\cdots$, f_{e_6} and uses an MLP layer to obtain the final emotional deformation ΔE.

Emotional Adaptation Module. The emotional adaptation module (EAM) consists of two multi-layer perceptrons (MLPs). As shown in Fig. 2(c), given the input feature x and the emotional token e_0, we extract the weight vector γ and the bias vector β using MLPs. They have the same dimension as the input x. With the channel-wise multiplication operation F_s and channel-wise addition, we obtain the output x'.

*Corresponding author.



(a) Emotion Mapper M (b) Emotion Deformation Network (c) Emotional Adaptation Module

Figure 2. More network architectures of our EAT model.

	Happy	Angry	Disgusted	Fear	Sad	Neutral	Surprised	Contempt	Average
Wav2Lip [14]	0.00	25.64	0.00	0.00	0.00	91.25	0.00	0.00	17.87
MakeItTalk [23]	0.00	25.64	0.00	0.00	0.00	75.00	0.00	0.00	15.23
AVCT [18]	0.83	25.64	0.00	0.00	0.00	69.38	0.00	10.08	15.64
EAMM [9]	23.33	84.48	9.40	0.00	0.00	98.13	94.02	72.27	49.85
Pretrain (Ours)	35.00	11.97	0.00	0.00	**49.17**	38.75	0.00	59.66	25.18
EAT (Ours)	**84.17**	**100.00**	**48.72**	**16.52**	**49.17**	**100.00**	**100.00**	**94.96**	**75.43**

Table 1. Quantitative evaluation of the emotion classification in the MEAD dataset.

Weight	PSNR↑	M/F-LMD↓	Sync↑	Acc_{emo}↑
w/o	21.49	2.27/2.46	8.02	**76**
EAT	**21.79**	**2.22/2.43**	**8.22**	67

Table 2. **Ablation study of EDN weight initialization.** The weight initialization of EDN with the A2ET encoder promotes the performance of EAT.

Parameter Efficiency Analysis. Our Deep Emotional Prompts, EDN and EAM only require about 7% of the parameters compared to the whole network. The emotion mapper, which generates deep emotional prompts for eight emotions, has most of the parameters. In addition, EDN and EAM consume less than 2%. These parameters are 13.9M. This is half of the emotional network of EAMM [9], which has 27.9M parameters.

B. Training and Testing Details

Training Details. We use the MEAD dataset and 8k emotional video clips from Voxceleb2 [5] with various facial expressions to learn the enhanced latent keypoints. We also use roughly 21k emotional images from AffectNet [13] to train emotional expression generation. Due to the lack of corresponding neutral faces, we generate neutral faces paired with emotional images by using Ganimation [15]. We train our EAT with Adam [10] with $\beta_1 = 0.5$ and $\beta_2 = 0.999$. The learning rate is set to 1.5×10^{-4} for A2ET and 2×10^{-4} for other modules. In the first stage, we train A2ET with only latent loss first to obtain a good initialization, and then we train it with full loss. To improve generalization, we use the Voxceleb2 and MEAD datasets, which contain about 225k video clips. In the second stage, we finetune efficient adaptation modules with only the MEAD dataset, which has about 10k video clips. We test our model on LRW [6] and MEAD [17] dataset.

Testing Details and Protocol. When testing LRW, the input is the first frame, and the transformation starts from the first frame. Therefore, the relative offsets of the latent keypoints are used. When testing MEAD, due to the variation in facial expressions, which is unrelated to the neutral source image, the predicted latent keypoints are used.

To ensure accurate evaluations, we crop and align [4] the faces before calculating these metrics: PSNR, SSIM, FID, M-LMD, and F-LMD. As for synchronization confidence, we preprocess the generated videos with reference to PC-AVS [22].

C. Additional Experimental Results

Additional baseline results. As shown in Figure 3, we

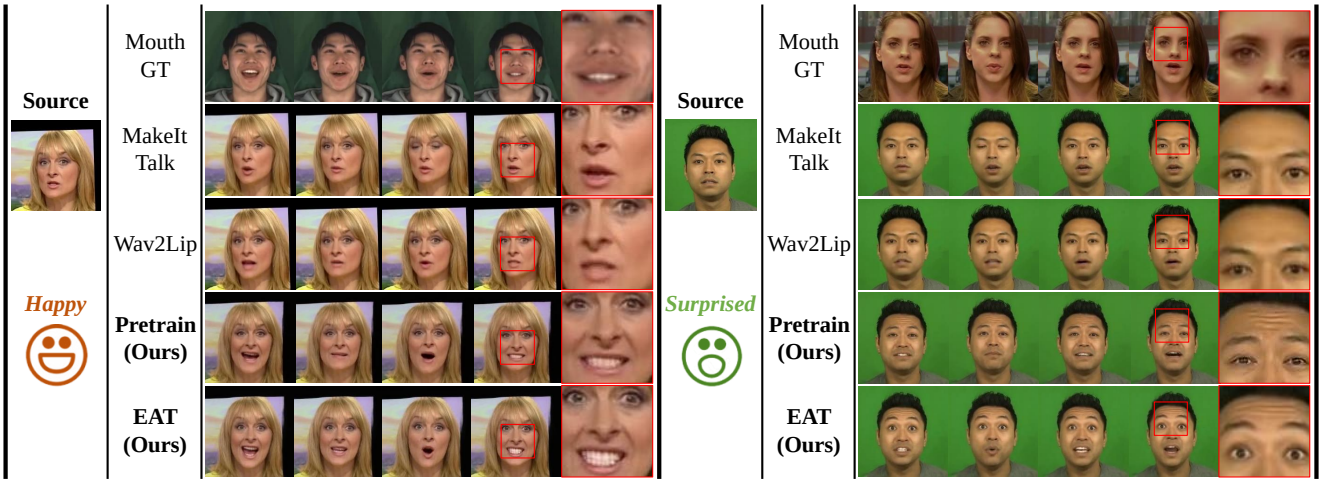


Figure 3. More Qualitative results. We compare with more baselines, such as MakeItTalk [23], Wav2Lip [14], and our pretrained model.

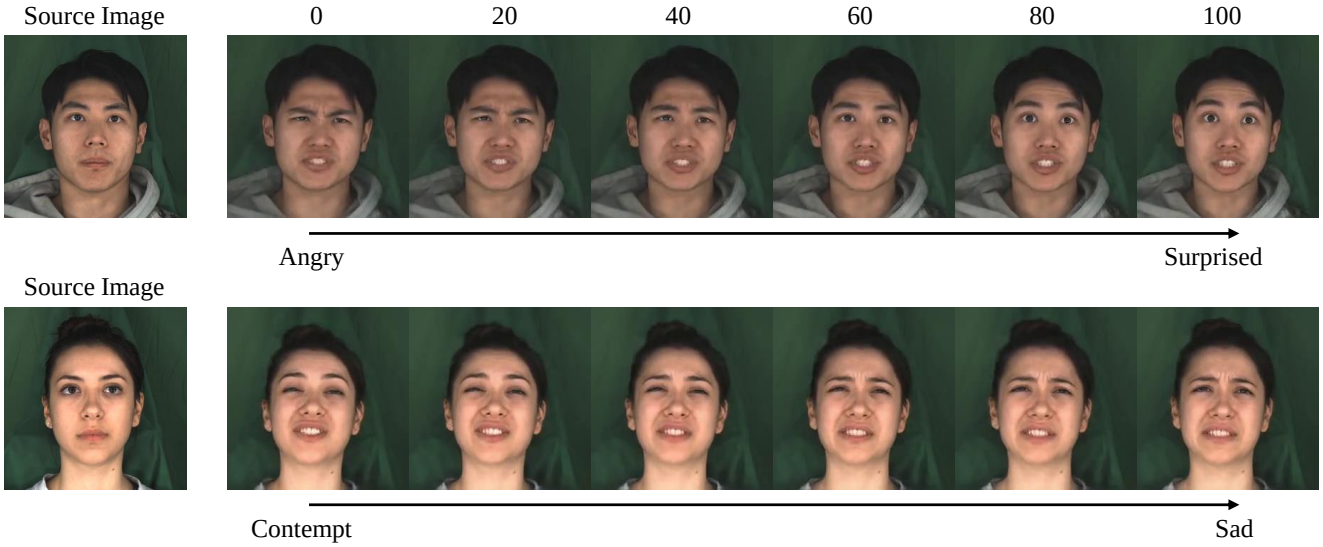


Figure 4. Emotion interpolation. The top row is the emotion interpolation results between $Angry$ and $Surprised$. The bottom row is the results between $Contempt$ and Sad. The neutral faces are from MEAD [17]

compare our EAT results with several baseline methods. Our results are more pleasant than those of MakeItTalk [23] and Wav2Lip [14], which do not consider emotional expression in talking heads. Additionally, our EAT achieves emotion control compared to the pretrained A2ET network. Videos are included in the supplementary material for reference.

Various emotional expressions To validate the diversity of emotional expressions generated by EAT, we present six different emotional results in Fig. 1. Compared to $Neutral$ emotion, emotional expressions result in different modifications to facial elements, such as lip corners, eyes, and brows. We present the quantitative results of emotion classification in Table 1. We notice that EAT works significantly better on $Happy$, Sad, $Disgusted$, and $Contempt$ than other methods. This is because our method can capture mouth details and these emotions can be more clearly reflected by the lips. As for $Neutral$, $Angry$, and $Surprised$, EAMM [9] performs well since these emotions are more apparent on the eyes and brows. And EAT can also achieve better performance in these emotions. However, all methods perform poorly on $Fear$ emotion. It may be because $Fear$ and $Surprise$ are similar and difficult to distinguish.

Emotion interpolation. We conduct emotional guidance interpolation on the MEAD test set to verify that the latent space learned by the emotion mapper is continuous, as Fig. 4 shows.

Additional ablation study. We conduct further ablation studies on the weight initialization of EDN, Our results, presented in Table 2, show that using the weight initialization of the A2ET encoder leads to quicker convergence and im-

Figure 5. Visualization on each component of EAT.

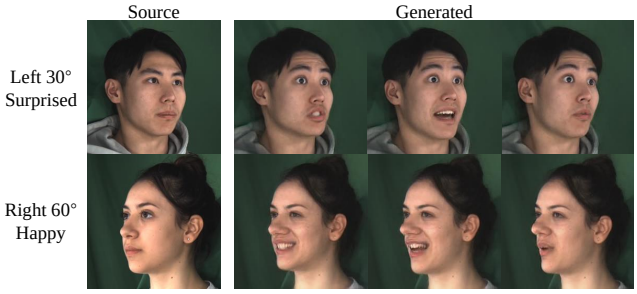


Figure 6. Visualization on the profile faces of MEAD.

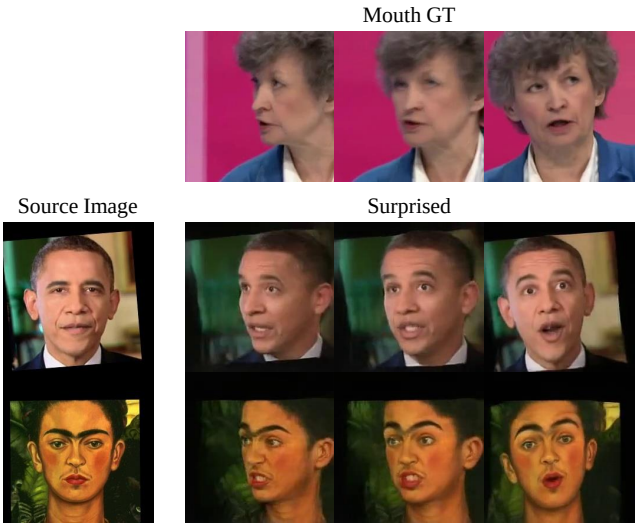


Figure 7. Failure case. The driving audio and poses are from the videos in the first row. The second and third rows display the generated results with $Surprised$ emotional guidance. Neutral faces are from MakeItTalk [23] and driving video is from LRW [6].

proved performance in terms of video quality and audio-visual synchronization.

Visual analysis on each component of EAT. To analyze the effect of each component of our model, we show the fear emotion results from (A), (B), and (C), with corresponding accuracy rates of 38.46%, 30.77%, and 15.38% respectively in Fig. 5. Deep emotional prompts help generate intense emotional expressions that deviate from the Ground Truth (GT). By incorporating EDM and EAM, we achieve greater fidelity toward the GT and higher image quality in terms of PSNR/SSIM. This is attributed to the learning capabilities of EDM and EAM for emotional data. However, it results in reduced emotion intensity and accuracy.

Visualization on the profile faces. To assess the ability of enhanced latent representation in 3D talking-head generation, as shown in Fig. 6, we visualize the talking-head frames generated from the profile faces of MEAD. We test the faces captured from left 30 degrees and right 60 degrees with $Suprised$ and $Happy$ emotions.

D. Limitations and Future Work.

While EAT is capable of generating emotional talking-head videos with emotional guidance, there are still some limitations. Firstly, the diversity of background and head pose in emotional training data can affect the generalizability of our EAT. As shown in Fig. 7, the wrinkles on the forehead are not obvious in these in-the-wild images. This issue could be addressed by more naturalistic and non-acted emotional data [20, 11, 2] and representations with the head prior, such as FLAME[21]. Secondly, effective guidance texts are required to achieve zero-shot generation. This may be due to the limited ability of models trained on image-text pairs to capture emotional expression, which could affect the performance of zero-shot learning. Thirdly, the eye region, such as eye blinks [16] and gaze direction [8], has not been considered in our work. Finally, the discrete emotion guidance limits the representation ability of our model. It needs to note that facial expressions are not always representative of the internal emotional state [1]. More refined theories of emotion, such as the valence-arousal model, may help generate more realistic emotions. We leave these problems for future work.

E. Ethical Considerations

Our research is intended for use in virtual human research and entertainment. However, there is a risk that the emotional talking-head generation algorithm could be abused. We strongly recommend that generated talking-head videos be labeled as "fake". On one hand, our work demonstrates that emotional talking-head generation

is technically feasible. On the other hand, fake video detection [7, 12] has attracted significant attention. We would be happy to assist in the development of related research.

References

[1] Lisa Feldman Barrett, Ralph Adolphs, Stacy Marsella, Aleix M Martinez, and Seth D Pollak. Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements. *Psychological science in the public interest*, 20(1):1–68, 2019. 4

[2] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42:335–359, 2008. 4

[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. 1

[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. 2

[5] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. VoxCeleb2: Deep Speaker Recognition. In *Interspeech 2018*, pages 1086–1090. ISCA, Sept. 2018. 2

[6] Joon Son Chung and Andrew Zisserman. Lip reading in the wild. In *Asian conference on computer vision*, pages 87–103. Springer, 2016. 2, 4

[7] Davide Cozzolino, Andreas Rossler, Justus Thies, Matthias Nießner, and Luisa Verdoliva. Id-reveal: Identity-aware deepfake video detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15108–15117, 2021. 5

[8] Michail Christos Doukas, Evangelos Ververas, Viktoriia Sharmanska, and Stefanos Zafeiriou. Free-headgan: Neural talking head synthesis with explicit gaze control. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 4

[9] Xinya Ji, Hang Zhou, Kaisiyuan Wang, Qianyi Wu, Wayne Wu, Feng Xu, and Xun Cao. EAMM: One-Shot Emotional Talking Face via Audio-Based Emotion-Aware Motion Model. In *Special Interest Group on Computer Graphics and Interactive Techniques Conference Proceedings*, pages 1–10, Vancouver BC Canada, Aug. 2022. ACM. 2, 3

[10] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *ICLR*, 2015. 2

[11] Jean Kossaifi, Robert Walecki, Yannis Panagakis, Jie Shen, Maximilian Schmitt, Fabien Ringeval, Jing Han, Vedhas Pandit, Antoine Toisoul, Björn Schuller, et al. Sewa db: A rich database for audio-visual emotion and sentiment research in the wild. *IEEE transactions on pattern analysis and machine intelligence*, 43(3):1022–1040, 2019. 4

[12] Momina Masood, Marriam Nawaz, Khalid Mahmood Malik, Ali Javed, and Aun Irtaza. Deepfakes generation and detection: State-of-the-art, open challenges, countermeasures, and way forward. *arXiv preprint arXiv:2103.00484*, 2021. 5

[13] A. Mollahosseini, B. Hasani, and M. H. Mahoor. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, PP(99):1–1, 2017. 2

[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. 2, 3

[15] Albert Pumarola, Antonio Agudo, Aleix M Martinez, Alberto Sanfeliu, and Francesc Moreno-Noguer. Ganimation: Anatomically-aware facial animation from a single image. In *Proceedings of the European conference on computer vision (ECCV)*, pages 818–833, 2018. 2

[16] Konstantinos Vougioukas, Stavros Petridis, and Maja Pantic. Realistic speech-driven facial animation with gans. *International Journal of Computer Vision*, 128:1398–1413, 2020. 4

[17] 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. Springer, 2020. 2, 3

[18] Suzhen Wang, Lincheng Li, Yu Ding, and Xin Yu. One-shot talking face generation from single-speaker audio-visual correlation learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 2531–2539, 2022. 2

[19] Ting-Chun Wang, Arun Mallya, and Ming-Yu Liu. One-shot free-view neural talking-head synthesis for video conferencing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10039–10049, 2021. 1

[20] Stefanos Zafeiriou, Dimitrios Kollias, Mihalis A Nicolaou, Athanasios Papaioannou, Guoying Zhao, and Irene Kotsia. Aff-wild: Valence and arousal 'in-the-wild'challenge. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2017 IEEE Conference on*, pages 1980–1987. IEEE, 2017. 4

[21] Yufeng Zheng, Victoria Fernández Abrevaya, Marcel C Bühler, Xu Chen, Michael J Black, and Otmar Hilliges. Im avatar: Implicit morphable head avatars from videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13545–13555, 2022. 4

[22] 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/CVF conference on computer vision and pattern recognition*, pages 4176–4186, 2021. 2

[23] Yang Zhou, Xintong Han, Eli Shechtman, Jose Echevarria, Evangelos Kalogerakis, and Dingzeyu Li. Makelttalk: speaker-aware talking-head animation. *ACM Transactions on Graphics (TOG)*, 39(6):1–15, 2020. 1, 2, 3, 4