FlowVQTalker: High-Quality Emotional Talking Face Generation through Normalizing Flow and Quantization

Supplementary Material

1. More Implementation Details

1.1. Preliminary of 3D Morphable Model

In our context, we mainly focus on the correlation between the driving audio a and the face expressions, with a strong commitment to preserving the identity and texture of the source image I. Therefore, it is imperative to disentangle these attributes from the image. Recent study [5] has led to promising results in single image 3D model [2] reconstruction, effectively separating the shape, texture, and lighting components of the input facial image. The decoupled shape can be further subdivided into expression coefficients, $\beta \in \mathbb{R}^{64}$, which control facial dynamics, and pose coefficients, $\rho \in \mathbb{R}^{6}$, which govern head movements. Inspired by this, we extract (β, ρ) from the input images and manipulate them to react to both audio input a and emotion label e.

1.2. Data Pre-processing Details

In the training process of FCG, we sample each training video at 25 fps, and re-crop the original videos to the resolution of 256×256 following [26]. We further leverage a 3D model reconstruction method [5] to extract expression coefficients and pose coefficients. Regarding the audio, we follow [22] to preprocess all the audio data to a uniform 16kHz sampling rate and compute mel-spectrograms using an FFT window size of 10 ms, a hop length of 200, and 80 Mel filter banks, resulting in a 16×80 mel-spectrogram feature for each frame. When pre-training the image encoder E_h, codebook $\mathcal{C}$ and image decoder D_h, we observed that they are sensitive to the quality of the training data and any background noise present. Consequently, we crop the original videos to the resolution of 512×512 and discard the noised background following [16].

2. More Implementation Details

2.1. Network Structure Details

FCG. In Fig. 3, we present an in-depth examination of the internal structures within the Flow-based Coefficient Generator (FCG), including the flow step f_K, the audio encoder E_a, and the emotion encoder E_e. A comprehensive explanation of the flow step f_K can be found in Sec. 3.2. The Transformer $\mathcal{F}$ within the coupling layer consists of 2 identical layers, each comprising 2 blocks: a local self-attention block with 8 heads, followed by a position-wise fully connected block. The audio Encoder E_a takes the 16×18-dimensional audio features as input and processes them through convolutional neural networks (CNN) followed by multi-layer perceptrons (MLP) to yield a 64-dimensional audio feature f_a. Meanwhile, the emotion encoder E_e is constructed with a multi-layer perceptron that maps the 9-dimensional emotion label e into a 32-dimensional emotion feature f_e.

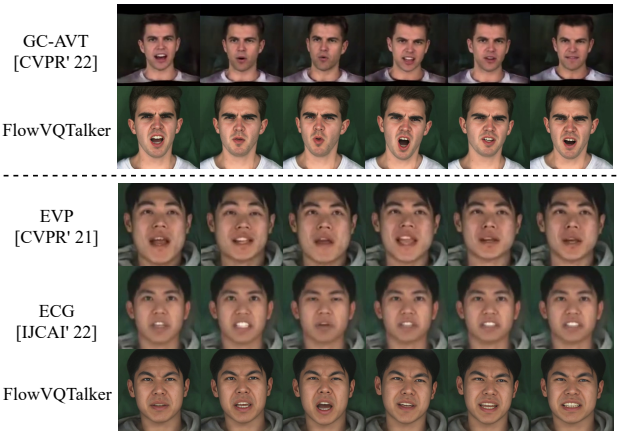
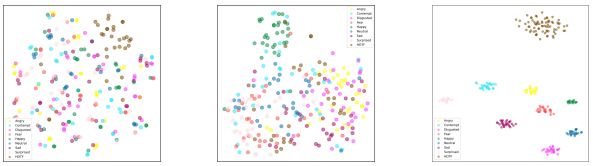


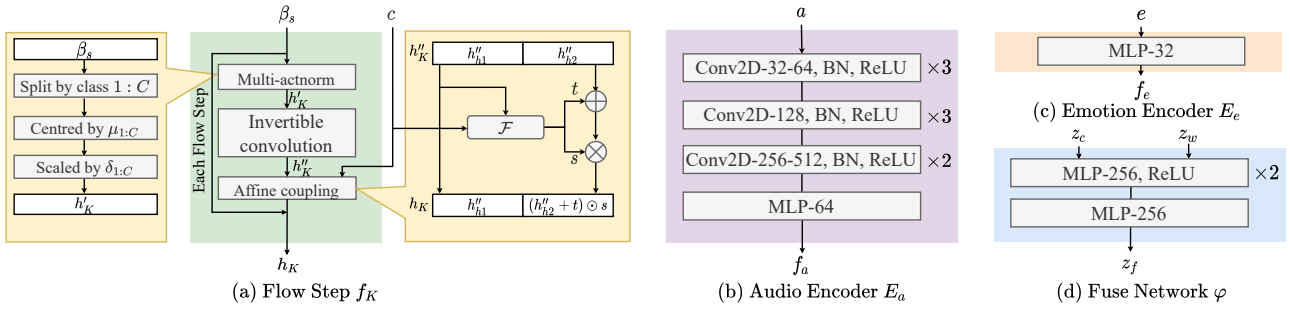
Figure 1. Comparison results with SOTA methods that have not released their codes and pretrained models.



(a) StyleTalk [20] (b) PD-FGC [30] (c) Ours

Figure 2. Visualization of latent space.

VQIG. We borrow the image encoder E_h, codebook $\mathcal{C}$ and image decoder D_h from VQGAN [6]. Please refer to the original paper for a comprehensive understanding of the network architecture. The warped image encoder E_w shares the same structure with E_h, with the distinction that it undergoes fine-tuning using the warped images generated by the motion descriptors Φ and the warping network W [24]. The fuse network φ is composed of a 3-layer multi-layer perceptron and combines z_c and z_W to form z_f.



(a) Flow Step f_K (b) Audio Encoder E_a (c) Emotion Encoder E_e (d) Fuse Network φ

Figure 3. Detailed architecture for different components in our FlowVQTalker.

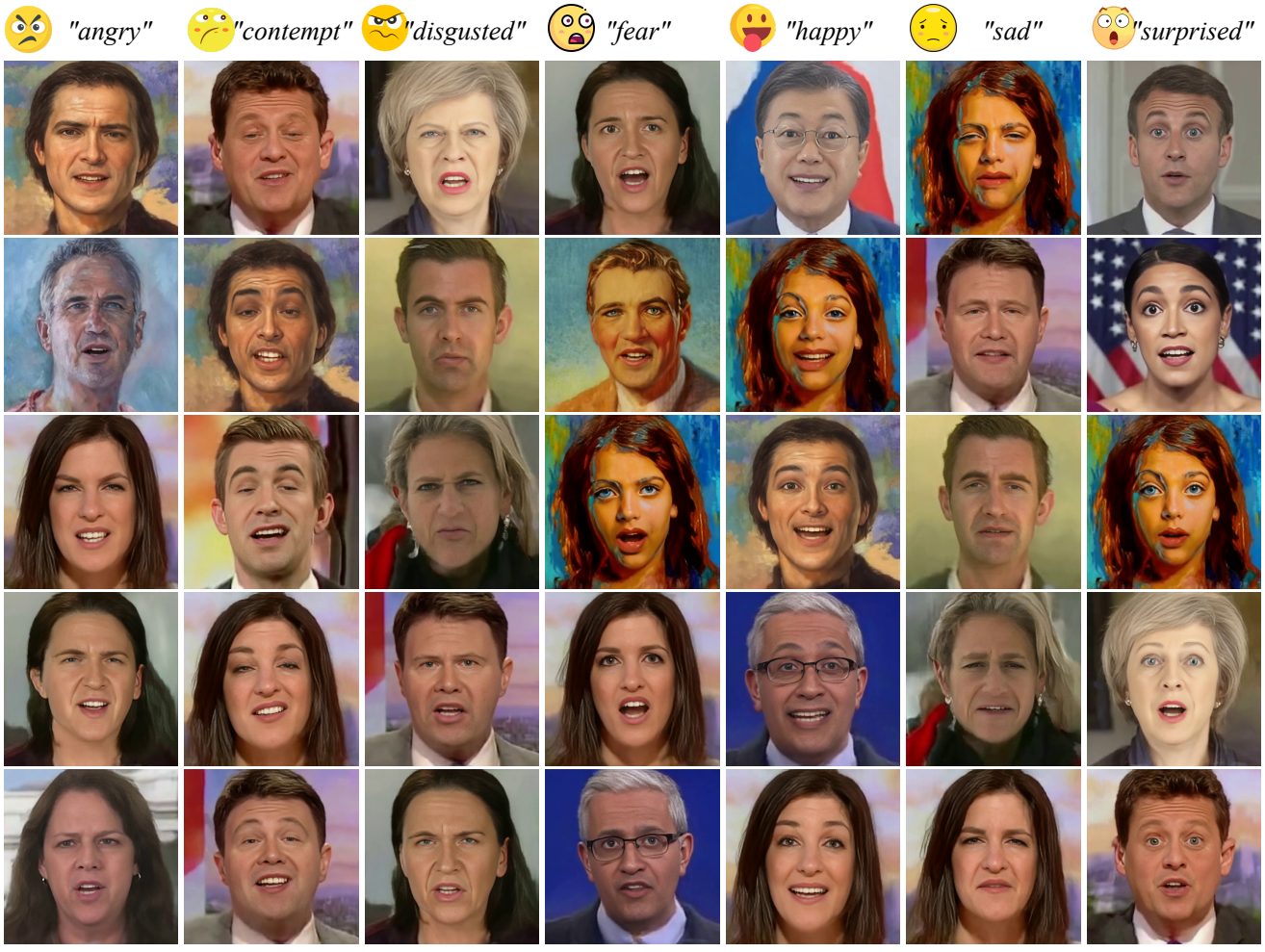


Figure 4. More results with various identities (including male, female, glass, paint) and various emotions (including angry, contempt, disgusted, fear, happy, sad, surprised).

Emotional label. Following previous works [7, 14, 27, 31], we utilize 8 discrete emotion labels (i.e., happy, angry, etc) ***annotated*** **by MEAD dataset** [31]. We further employ a more fine-grained emotion label by extracting the expression coefficients of emotion reference image using 3DMM ***only*** for emotion transfer. As observed and verified

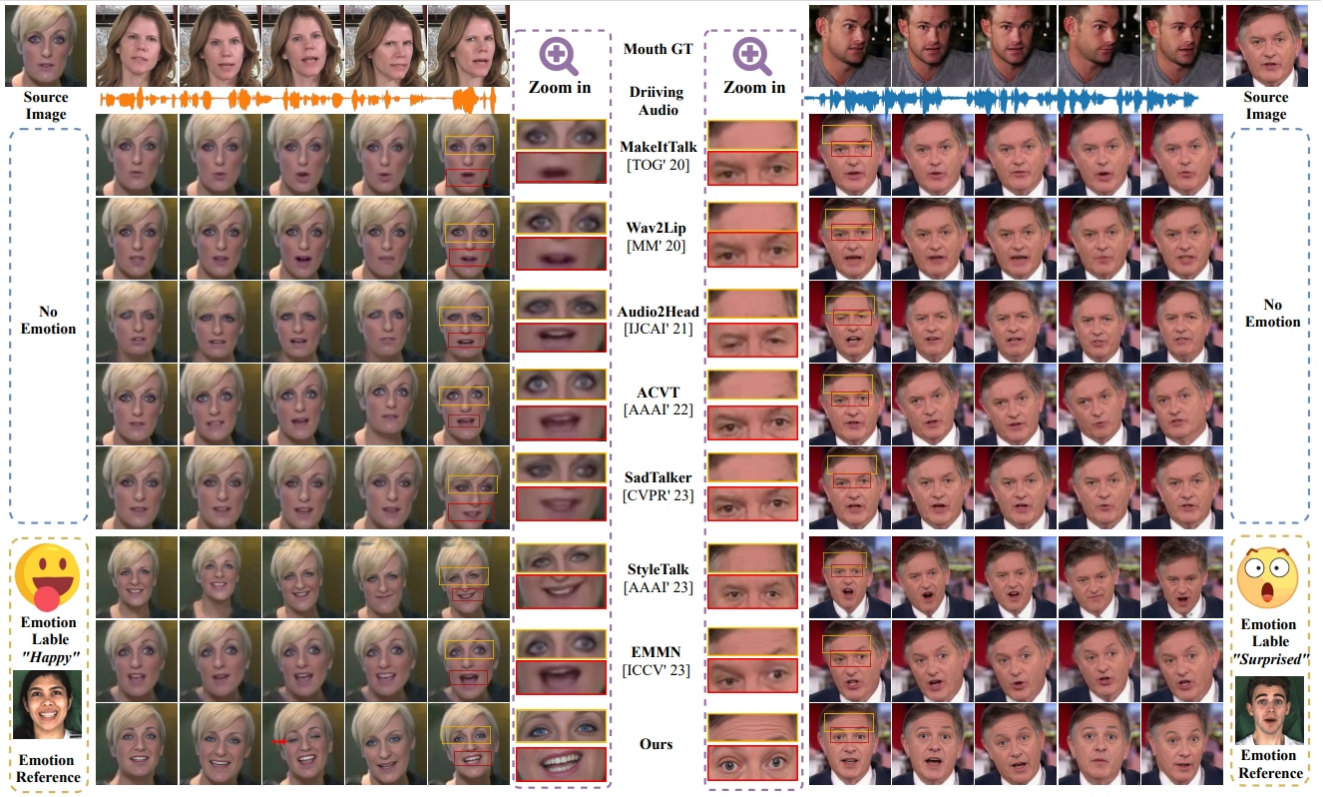


Figure 5. Additional qualitative results, which are supplement to Fig. 4 of the main paper.

by [7, 20, 24, 42] is actually robust to identity, expression, age and gender of input images, and thus [5] is competent to extract the emotion label. Although the emotion has a few levels in speaker, most of them can be grouped into 8 discrete labels described above. Therefore, when evaluating emotion accuracy, we first generate an emotional video conditioned on the input audio, image and emotion label (e.g., input emotion label: happy). Then we employ an emotion classification network [21] to predict the emotion category of the generated video. If the predicted emotion category aligns with the input emotion label (e.g., predicted emotion: happy; input emotion label: happy), we consider that we generate a video with the correct emotion, and vice versa (e.g., predicted emotion: sad; input emotion label: happy) to calculate the emotion accuracy. As for the influence of emotional audio on expression, we address the concern by extracting acoustic features from audio and training our audio encoder to prioritize content-related information over emotion-related information.

2.2. Training and Evaluation Details

In the configuration of the flow-based coefficient generator, we specify the following parameters: the number of flow steps K is set to 16; the previous frame length τ, within the context c, is fixed at 10; the value of ν in Eq. (9) is chosen as 50; the dropout rate is established at 0.7. We assign weights of 1 for $\mathcal{L}_{\text{exp}}$ in Eq. (10) and 1e-3 for $\mathcal{L}_{\text{con}}$ in Eq. (11). Furthermore, the FCG and VQIG models are trained with learning rates of 1e-4 and 2e-4, respectively. To avoid the impact of different face cropping methods on the metrics [28], we employ face cropping and alignment techniques [17] for calculating metrics such as: SSIM, FID, M-LMD, F-LMD and CPBD. Additionally, we utilize the cropping method in [44] for computing $\text{Sync}_{\text{conf}}$.

3. More Experiment Results

3.1. Various identities with various expressions

Fig. 4 showcases additional results featuring a diverse range of identities and expressions. Our method exhibits robustness across a spectrum of identities, encompassing male and female, individuals wearing glasses and paints.

3.2. More Comparison Results

Additional Comparison Results with SOTAs. Apart from the state-of-the-art (SOTA) methods discussed in the main paper, we extend our comparative analysis to include both emotion-agnostic talking face generation methods: MakeItTalk [47], Audio2Head [34], AVCT [35], and SadTalker [42], as well as emotional talking face generation

Method	MEAD [31]						HDTF [43]				
	SSIM↑	FID↓	M-LMD↓ / F-LMD↓	$\text{Sync}_{\text{conf}}$ ↑	CPBD↑	Acc_{emo} ↑	SSIM↑	FID↓	M-LMD↓ / F-LMD↓	$\text{Sync}_{\text{conf}}$ ↑	CPBD↑
MakeItTalk [48]	0.634	25.618	2.504 / 2.342	4.835	0.112	17.62%	0.707	16.582	2.264 / 2.053	4.546	0.137
Audio2Head [34]	0.620	28.700	2.477 / 2.497	5.575	0.123	15.87%	0.706	17.312	1.766 / 2.075	5.071	0.137
AVCT [35]	0.612	21.341	2.664 / 2.857	5.075	0.136	14.63%	0.665	17.805	2.152 / 2.372	5.346	0.134
SadTalker [42]	0.631	20.444	2.143 / 2.322	5.726	0.148	14.58%	**0.720**	**13.182**	1.716 / **1.815**	**7.368**	0.150
StyleTalk [20]	**0.758**	21.950	2.188 / 2.126	2.946	0.139	67.35%	0.690	15.646	2.102 / 2.096	2.418	0.166
FlowVQTalker	0.689	**16.553**	**1.939 / 2.061**	**5.901**	**0.181**	**71.53%**	0.708	15.165	**1.643** / 1.958	6.766	**0.268**
GT	1.000	0.000	0.000 / 0.000	6.733	0.161	81.68%	1.000	0.000	0.000 / 0.000	7.728	0.238

Table 1. Quantitative comparisons with state-of-the-art methods. We test each method on MEAD and HDTF datasets, and the best scores in each metric are highlighted in bold. The signages " ↑ " and " ↓ " indicate higher and lower metric values for better results, respectively.

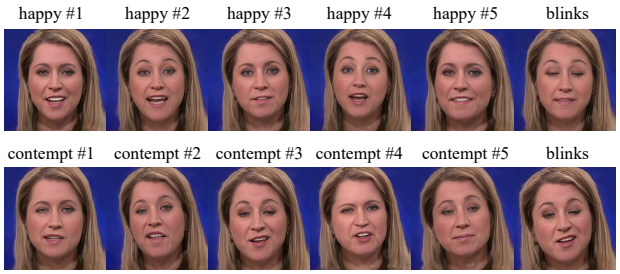


Figure 6. Our results with diverse and synchronous facial dynamic, including expressions, blinks, poses even in the case of identical inputs.

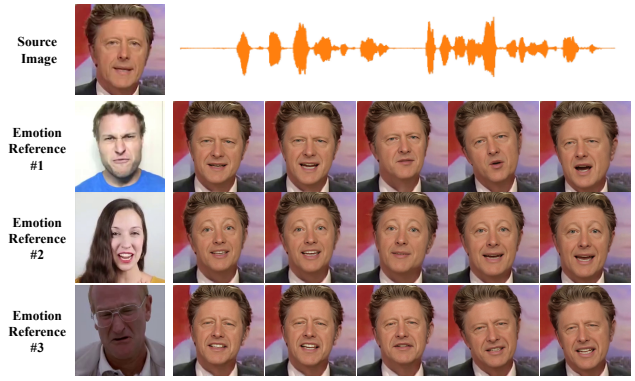


Figure 7. Emotion transfer results. Given a source image and an emotion reference, our method is capable of achieving emotion transfer.

methods: StyleTalk [20]. The comprehensive qualitative and quantitative results can be found in Fig. 5 and Tab. 1, serving as a supplement to the previously presented data in Fig. 4 and Tab. 1 of the main paper. Furthermore, Fig. 1 provides a comparison with SOTA methods that have not made their code and pretrained models publicly available: GC-AVT [18], EVP [14], and ECG [27]. Our approach outperforms these methods, demonstrating superior lip-synchronization, expressive facial emotion dynamics, and high-definition textures. Moreover, we perform an comparison with EmoGen [8] displayed in Fig. 11. Our method generates more realistic expressions with higher image quality, whereas EmoGen exhibits severe identity loss and prominent boundaries (highlighted by red box).

Metric/Method	MakeItTalk	Audio2Head	AVCT	SadTalker	Ours	GT
Diversity ↑	0.616	1.410	**2.247**	1.527	1.911	1.846
CCA ↑	0.739	0.771	0.793	0.775	**0.801**	-

Table 2. Quantitative comparison for poses. Diversity is calculated using the standard deviation of the 6-D pose and CCA [11] evaluates the correlation between generated poses and GT.

Method/Score	SSIM ↑	FID ↓	M-LMD ↓	F-LMD ↓	$\text{Sync}_{\text{conf}}$ ↑	CPBD ↑
w/o data dropout	0.682	21.374	2.446	2.513	3.447	0.168
Baseline	0.652	41.493	2.098	2.165	5.147	0.133
B+W	0.675	27.431	2.059	2.253	5.575	0.147
B+W+φ	0.672	20.232	2.084	2.355	5.837	0.146
B+W+φ+M	0.680	20.936	2.014	2.101	5.947	0.155
Full Model	**0.689**	**16.553**	**1.939**	**2.061**	**5.901**	**0.181**

Table 3. Results for ablation study on MEAD dataset.

Latent Space Comparison. We conduct a comparison with StyleTalk [20] and PD-FGC [30], both of which delve into the exploration of latent space for emotions. In Fig. 2, we visualize the latent space using t-distributed stochastic neighbor embedding (t-SNE) [29]. Notably, our method demonstrates superior clustering in the latent space compared to the comparison methods [20, 30].

3.3. Ablation Study

Ablation on FCG. Thanks to the random sampling capability of our proposed flow-based coefficient generator, we can generate multiple talking face videos even when the input remains consistent. Fig. 6 demonstrates two examples featuring diverse facial emotion dynamics, encompassing different expressions, blinks, and head poses. Notably, "happy#1"-"happy#5" and "contempt#1"-"contempt#5", even though they express the same emotion, are different in the speaking style, emotion intensity, facial details (such as eye width and mouth corner

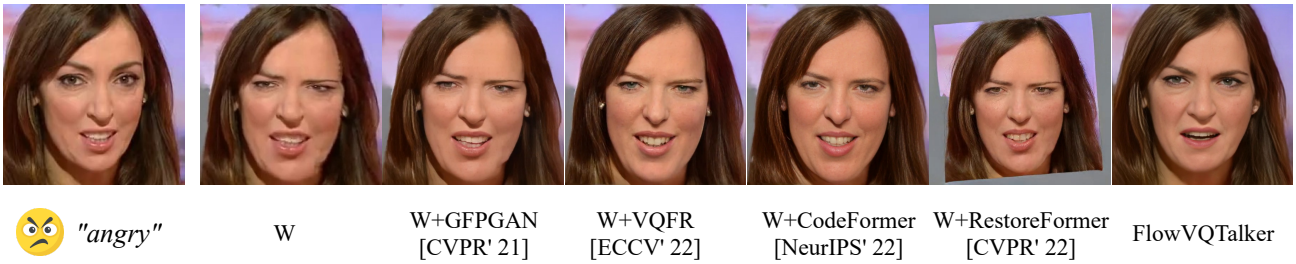


Figure 8. Comparison with SOTA face restoration methods.

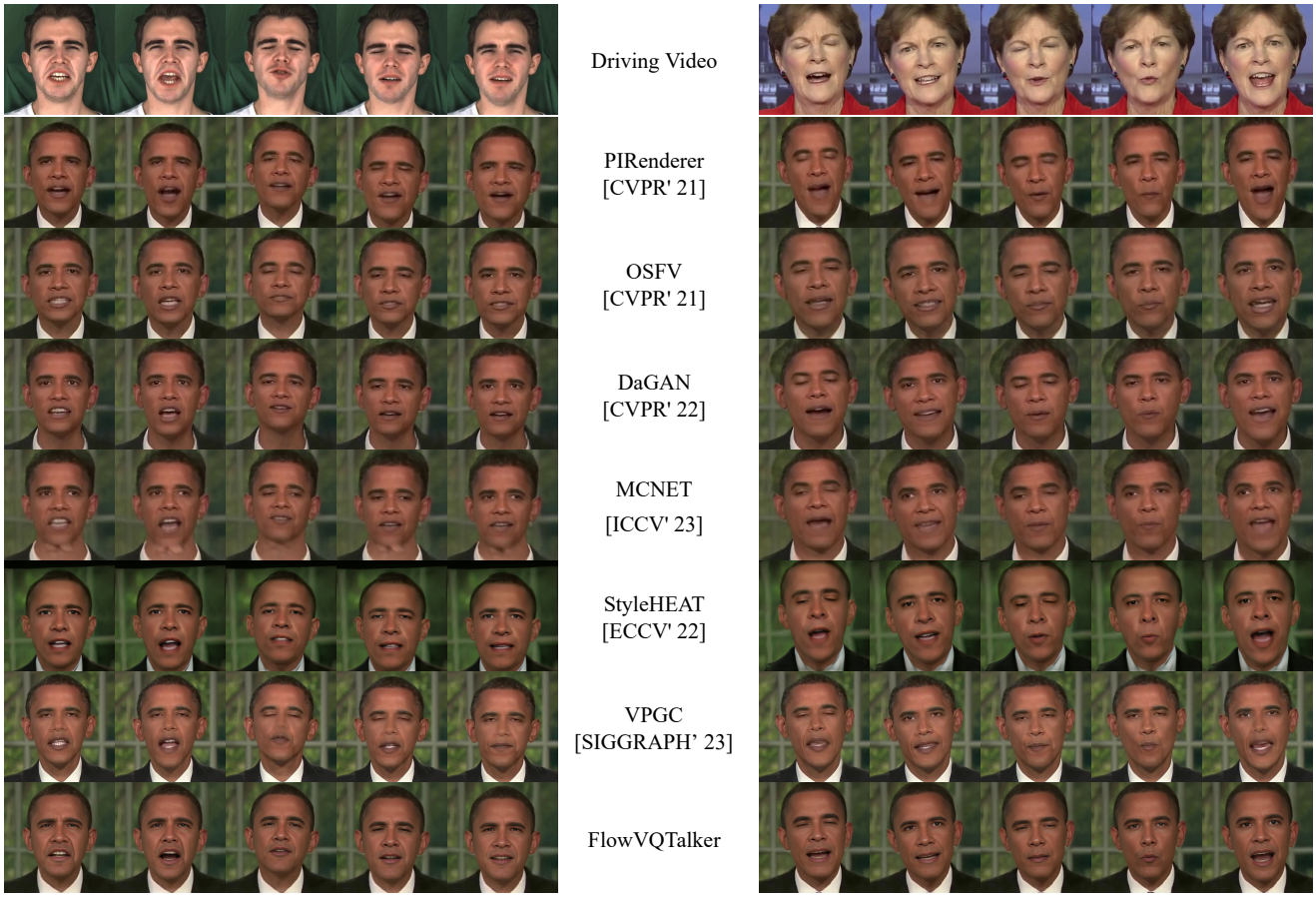


Figure 9. Comparison with SOTA face reenactment methods.

angle), and more. Furthermore, given the bijective nature of our proposed FCG, we support emotion transfer by providing an emotion reference. As depicted in Fig. 7, our FlowVQTalker adeptly replicates the expression of a specified reference and generates synchronized mouth shapes corresponding to the audio.

Fig. 5 of the main paper and Fig. 6 provide visualizations of a wide range of head poses generated by our PoseFlow model. To facilitate a more intuitive quantitative comparison, we compute the standard deviation of the 6-dimensional head pose parameters to gauge the diversity of the generated head motions. Additionally, we employ Canonical Correlation Analysis (CCA) [11] to assess the correlation between the generated poses and the ground truth. The results presented in Tab. 2 demonstrate the superior performance of our method in terms of diversity and similarity.

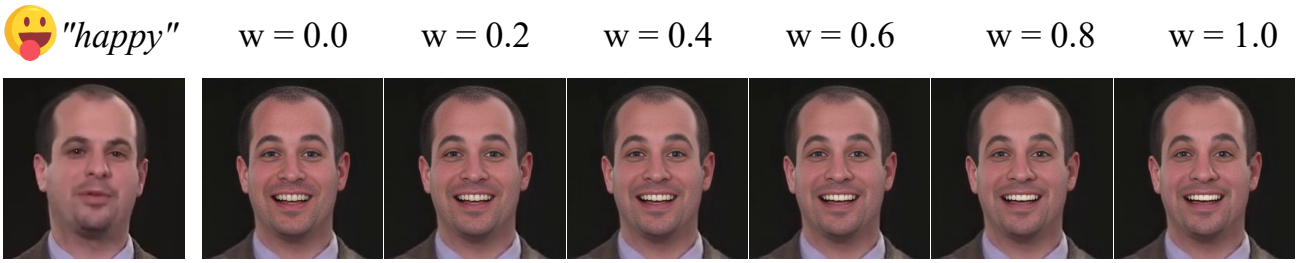


Figure 10. Trade-off between fidelity (SSIM) and quality(CPBD).

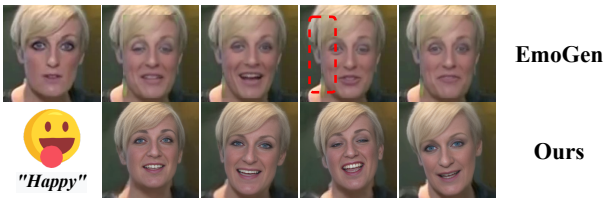


Figure 11. Comparison with EmoGen

Ablation on VQIG. We remove ADAIN and the corresponding coefficient input from VQIG to further assess its contribution. We observe that I_w has obvious artifacts, inadequate movement and inaccurate expression. Since w/o ADAIN only takes I_w as input and no additional motion descriptor, it can only rectify the artifacts. Subsequently, we further offer coefficients via ADAIN at both feature and image levels, leading to enhanced facial expressiveness. Consequently, the fine-tuning network can be viewed as a combination of a super-resolution network and a motion-complementary network.

In addition to qualitative results presented in Fig. 7 of the main paper, we provide quantitative results in Tab. 3, confirming the effectiveness of each module within the ExpFlow and VQIG. Additionally, we attempt SOTA face restoration methods (i.e., GFPGAN [38], VQFR [9], CodeFormer [45], RestoreFormer [40]), to complement high-quality details lost in the warped image I_w. We consider the following variants: (1) Only W: Utilizing only W to obtain the final results. (2) W+GFPGAN: Employing GFPGAN to enhance the quality of results generated in (1). (3) W+VQFR. (3) W+CodeFormer. (3) W+RestoreFormer. As demonstrated in Fig. 8, our FlowVQTalker stands out as the sole method capable of generating expressive facial expressions with the desired emotion while preserving the identity and texture of the source image.

By extracting the coefficients of a given driving video, our VQIG enables practical face reenactment. In this context, we conduct a comparison with state-of-the-art face reenactment methods, which include PIRenderer [24], OSFV [36], DaGAN [13], MCNET [12], StyleHEAT [41], and VPGC [33], where StyleHEAT and VPGC are capable of generating high-resolution results, and VPGC is a person-specific model. Given that the compared methods are not specifically trained on emotional datasets, we conduct comparisons using videos with and without emotion, the results of which are presented in Fig. 9. Our method excels in generating emotion-aware textures and delivers the best performance in terms of face reenactment.

While the fuse network φ and MHCA [40] are designed to preserve the identity information of the source image, there is often a trade-off between quality and fidelity [45], both of which are vital aspects of talking face generation. To this end, we enable a user-friendly control of the trade-off by introducing the controllable feature transformation (CFT) module using w as inspired by [45]. The results, depicted in Fig. 10, illustrate that a higher value of w results in higher fidelity (SSIM) but lower quality (CPBD), and vice versa. Furthermore, we conduct a comparison experiment with SOTA methods in terms of identity preservation ability. To this end, we introduce cosine similarities (CSIM) of the identity features extracted by a face recognition network F_r [4] as a quantitative metric. The results are reported in Tab. 4. Remarkably, only two emotion-agnostic methods (i.e., MakeItTalk and Wav2Lip) outperform our FlowVQTalker and we regard it as reasonable. One the one hand, Wav2Lip excels by exclusively modifying the mouth regions while leaving other facial components unchanged, thereby achieving the highest CSIM score. On the other hand, there exists an empirical trade-off between emotional expressions and identity preservation. Various expressions significantly impact the extraction of identity features by F_r, leading to a reduction in CSIM. It is noteworthy that, within the context of generating emotionally expressive talking faces, our FlowVQTalker outperforms all alternative methods.

MakeItTalk [48]	Wav2Lip [22]	Audio2Head [34]	PC-AVS [44]	EAMM [15]	StyleTalk [20]	PD-FGC [30]	EAT [7]	FlowVQTalker
0.791	**0.856**	0.664	0.636	0.473	0.754	0.469	0.753	0.768

Table 4. Comparison of identity preservation using CSIM.

3.4. User Study Setting

In order to avoid excessively long testing times that could impact the participants' judgment, we opt to select only two state-of-the-art emotion-agnostic and emotional methods for comparison in each category, rather than evaluating all 14 methods (comprising 7 emotion-agnostic methods, 6 emotional methods, and ground truth), which would entail the generation of 280 videos (14 methods x 20 samples) and consume a significant amount of time.

4. Discussion

4.1. Limitation and Future Work

Despite achieving satisfactory performance, our method still has the following limitations. Firstly, the codebook $\mathcal{C}$ in our VQIG is trained on cropped faces without background information, which means that the backgrounds in the results can only maintain the same level of sharpness as the input image. Similar to approaches like VPNQ [32] and VPGC [33], retraining the codebook using a large, high-quality face dataset with background information could address this limitation. However, this process demands significantly more computational resources, is prone to unstable training, and can be challenging to converge. Alternatively, for simplicity, a background enhancement technique [37, 39] can be applied to enhance the background of the input image, resulting in high-definition (HD) outputs. Secondly, FlowVQTalker does not currently support text-driven emotion generation. In future work, we plan to explore the integration of large language models, such as CLIP [23], to extract features from text and incorporate them into the context, potentially enabling zero-shot emotion editing [7, 19]. Furthermore, the challenge of interpolating emotions, like sad to happy, arises due to the strong dependence of the generated expressions on the sampled codes from the modeled distribution and the invertible mapping of the ExpFlow, as the interpolated codes are not within the modeled distribution and cannot be mapped in an invertible manner to achieve the desired results. Thirdly, FlowVQTalker faces challenges in generating real-time talking-head videos. On the one hand, the inherent limitations of normalizing flow [25], which involves generating each frame of motion and requires multiple flow steps, can impede inference efficiency. On the other hand, VQIG generates higher-resolution (512×512) and higher-quality images compared to previous methods (typically 224×224 or 256×256), leading to increased inference time. Lastly, when faced with distorted face images from which coefficients cannot be extracted, mode collapse occurs, as is a common issue with 3DMM-based methods.

4.2. Ethical considerations.

Our approach is designed to create emotionally expressive talking avatars with a high degree of realism. While this technology offers significant advantages for avatar-related research and enhances various aspects of people's leisure activities, such as video conferencing and virtual reality, it also carries the risk of potential misuse, which could have adverse societal consequences. In light of this, we recommend attaching a watermark to the generated video, allowing users to readily differentiate them. Concurrently, we are open to collaborating with deepfake detection research efforts [1, 3, 10, 46], offering our generated videos to bolster their performance and mitigate any potential negative repercussions.

References

[1] Relja Arandjelovic and Andrew Zisserman. Look, listen and learn. In *Proceedings of the IEEE international conference on computer vision*, pages 609–617, 2017. 7

[2] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3d faces. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, pages 187–194, 1999. 1

[3] Lucy Chai, David Bau, Ser-Nam Lim, and Phillip Isola. What makes fake images detectable? understanding properties that generalize. *Lecture Notes in Computer Science*, 2020. 7

[4] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4690–4699, 2019. 6

[5] Yu Deng, Jiaolong Yang, Sicheng Xu, Dong Chen, Yunde Jia, and Xin Tong. Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 0–0, 2019. 1, 3

[6] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 1

[7] Yuan Gan, Zongxin Yang, Xihang Yue, Lingyun Sun, and Yi Yang. Efficient emotional adaptation for audio-driven talking-head generation. In *Proceedings of the IEEE/CVF*

International Conference on Computer Vision, pages 22634–22645, 2023. 2, 3, 7

[8] Sahil Goyal, Sarthak Bhagat, Shagun Uppal, Hitkul Jangra, Yi Yu, Yifang Yin, and Rajiv Ratn Shah. Emotionally enhanced talking face generation. In *Proceedings of the 1st International Workshop on Multimedia Content Generation and Evaluation: New Methods and Practice*, pages 81–90, 2023. 4

[9] Yuchao Gu, Xintao Wang, Liangbin Xie, Chao Dong, Gen Li, Ying Shan, and Ming-Ming Cheng. Vqfr: Blind face restoration with vector-quantized dictionary and parallel decoder. In *European Conference on Computer Vision*, pages 126–143. Springer, 2022. 6

[10] David Guera and Edward J. Delp. Deepfake video detection using recurrent neural networks. *Advanced Video and Signal Based Surveillance*, 2018. 7

[11] GregoryR. Hancock, RalphO. Mueller, and LauraM. Stapleton. *Canonical Correlation Analysis*, page 45–56. 2010. 4, 5

[12] Fa-Ting Hong and Dan Xu. Implicit identity representation conditioned memory compensation network for talking head video generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23062–23072, 2023. 6

[13] Fa-Ting Hong, Longhao Zhang, Li Shen, and Dan Xu. Depth-aware generative adversarial network for talking head video generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3397–3406, 2022. 6

[14] 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*, pages 14080–14089, 2021. 2, 4

[15] 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 *ACM SIGGRAPH 2022 Conference Proceedings*, pages 1–10, 2022. 7

[16] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 1

[17] Chen Lele, K Maddox Ross, Duan Zhiyao, and Xu Chenliang. Hierarchical cross-modal talking face generation with dynamic pixel-wise loss. *IEEE Conference Proceedings*, 2019. 3

[18] Borong Liang, Yan Pan, Zhizhi Guo, Hang Zhou, Zhibin Hong, Xiaoguang Han, Junyu Han, Jingtuo Liu, Errui Ding, and Jingdong Wang. Expressive talking head generation with granular audio-visual control. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3387–3396, 2022. 4

[19] Yifeng Ma, Suzhen Wang, Yu Ding, Bowen Ma, Tangjie Lv, Changjie Fan, Zhipeng Hu, Zhidong Deng, and Xin Yu. Talkclip: Talking head generation with text-guided expressive speaking styles. *arXiv preprint arXiv:2304.00334*, 2023. 7

[20] Yifeng Ma, Suzhen Wang, Zhipeng Hu, Changjie Fan, Tangjie Lv, Yu Ding, Zhidong Deng, and Xin Yu. Styletalk: One-shot talking head generation with controllable speaking styles. *arXiv preprint arXiv:2301.01081*, 2023. 1, 3, 4, 7

[21] Debin Meng, Xiaojiang Peng, Kai Wang, and Yu Qiao. Frame attention networks for facial expression recognition in videos. In *2019 IEEE international conference on image processing (ICIP)*, pages 3866–3870. IEEE, 2019. 3

[22] 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, 7

[23] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 7

[24] Yurui Ren, Ge Li, Yuanqi Chen, Thomas H Li, and Shan Liu. Pirenderer: Controllable portrait image generation via semantic neural rendering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13759–13768, 2021. 1, 3, 6

[25] Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *International conference on machine learning*, pages 1530–1538. PMLR, 2015. 7

[26] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. First order motion model for image animation. *neural information processing systems*, 2019. 1

[27] Sanjana Sinha, Sandika Biswas, Ravindra Yadav, and Brojeshwar Bhowmick. Emotion-controllable generalized talking face generation. 2, 4

[28] Shuai Tan, Bin Ji, and Ye Pan. Emmn: Emotional motion memory network for audio-driven emotional talking face generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22146–22156, 2023. 3

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

[30] Duomin Wang, Yu Deng, Zixin Yin, Heung-Yeung Shum, and Baoyuan Wang. Progressive disentangled representation learning for fine-grained controllable talking head synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17979–17989, 2023. 1, 4, 7

[31] 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 *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXI*, pages 700–717. Springer, 2020. 2, 4

[32] Kaisiyuan Wang, Changcheng Liang, Hang Zhou, Jiaxiang Tang, Qianyi Wu, Dongliang He, Zhibin Hong, Jingtuo Liu,

Errui Ding, Ziwei Liu, and Jingdong Wang. Robust video portrait reenactment via personalized representation quantization. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(2):2564–2572, 2023. 7

[33] Kaisiyuan Wang, Hang Zhou, Qianyi Wu, Jiaxiang Tang, Zhiliang Xu, Borong Liang, Tianshu Hu, Errui Ding, Jingtuo Liu, Ziwei Liu, et al. Efficient video portrait reenactment via grid-based codebook. In *ACM SIGGRAPH 2023 Conference Proceedings*, pages 1–9, 2023. 6, 7

[34] S Wang, L Li, Y Ding, C Fan, and X Yu. Audio2head: Audio-driven one-shot talking-head generation with natural head motion. In *International Joint Conference on Artificial Intelligence*. IJCAI, 2021. 3, 4, 7

[35] 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*, pages 2531–2539, 2022. 3, 4

[36] 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. 6

[37] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. Esrgan: Enhanced super-resolution generative adversarial networks. In *Proceedings of the European conference on computer vision (ECCV) workshops*, pages 0–0, 2018. 7

[38] Xintao Wang, Yu Li, Honglun Zhang, and Ying Shan. Towards real-world blind face restoration with generative facial prior. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9168–9178, 2021. 6

[39] Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1905–1914, 2021. 7

[40] Zhouxia Wang, Jiawei Zhang, Runjian Chen, Wenping Wang, and Ping Luo. Restoreformer: High-quality blind face restoration from undegraded key-value pairs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17512–17521, 2022. 6

[41] Fei Yin, Yong Zhang, Xiaodong Cun, Mingdeng Cao, Yanbo Fan, Xuan Wang, Qingyan Bai, Baoyuan Wu, Jue Wang, and Yujiu Yang. Styleheat: One-shot high-resolution editable talking face generation via pre-trained stylegan. In *European conference on computer vision*, pages 85–101. Springer, 2022. 6

[42] Wenxuan Zhang, Xiaodong Cun, Xuan Wang, Yong Zhang, Xi Shen, Yu Guo, Ying Shan, and Fei Wang. Sadtalker: Learning realistic 3d motion coefficients for stylized audio-driven single image talking face animation. *arXiv preprint arXiv:2211.12194*, 2022. 3, 4

[43] Zhimeng Zhang, Lincheng Li, Yu Ding, and Changjie Fan. Flow-guided one-shot talking face generation with a high-resolution audio-visual dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3661–3670, 2021. 4

[44] 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. 3, 7

[45] Shangchen Zhou, Kelvin Chan, Chongyi Li, and Chen Change Loy. Towards robust blind face restoration with codebook lookup transformer. *Advances in Neural Information Processing Systems*, 35:30599–30611, 2022. 6

[46] Yipin Zhou and Ser-Nam Lim. Joint audio-visual deepfake detection. *International Conference on Computer Vision*, 2021. 7

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

[48] 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. 4, 7