EmoTalker Supplementary Materials

1 Compare with emotional approach

All the previous emotional talking head approaches rely on driving frames to generate facial expressions that ignore the emotional signals conveyed by the audio itself. Our motivation is to model facial expressions from input audio without additional emotion information. Therefore, our setting stands apart as it exclusively utilizes audio input and doesn't require extra emotional driving frames. Therefore, it is unfair to compare with those methods.

However, we enable a fair comparison with previous emotion-based works with two scenarios. Firstly, we fit the previous method to our setting by making all the driving frames to be the neutral identity image. Second, we adopt our model for the previous setting by taking both audio and emotional driving frames as input to generate videos. The comparison result is shown in Table 1. We compare with EAMM [4] since EVP[1] provides only two separate pre-trained models of two target persons. As Table 1 shows, our method achieves comparable performance when input with driving frames with EAMM [4]. And when w/o input driving frames, our method outperforms it in each aspect, which indicates our model captures the emotional signals from audio without the help of external emotional information.

Method	**w/ driving frames**	**Texture Quality**		**Landmark Alignment**		**Lip Sync**	**Emotion**
		SSIM ↑	CPBD ↑	F-LMD ↓	M-LMD ↓	Sync_{conf} ↑	Emo_{Acc} ↑
EAMM [13]	✓	0.665	0.316	2.541	3.672	1.837	0.718
EmoTalker (ours)	✓	0.695	0.372	2.363	3.842	2.538	0.762
EAMM [13]	×	0.660	0.307	7.985	5.133	1.605	0.183
EmoTalker (ours)	×	0.705	0.362	4.348	5.139	2.304	0.425

Table 1: Compare with emotional talking head work EAMM [4]. In the upper section, we adapt our model to the previous setting by taking both audio and emotional driving frames as input to generate videos. In the lower section, we fit EAMM to our setting by making all the driving frames to be the neutral identity image.

2 Qualitative Results

Please refer to our webpage for more qualitative results, including SoTA comparison, diverse emotion-aware synthesized videos, rendered 3D face videos, free control videos

[1] https://github.com/jixinya/EVP

by driven (pose, expression and lip) frames and higher-resolution videos enhanced by GFPGAN [19].

3D Faces. Our model first generates 3DMM coefficients with driven audio and obtains the identity information extracted from a reference image. Figure 3 illustrates the textured 3DMM faces with various identities. For a live demonstration of the dynamic video, please visit our webpage.

Diverse Emotion Aware Videos. Our webpage showcases generated examples using a neutral identity reference, with the same speech content but varying audio fluctuations. Previous methods struggle to differentiate emotional signals from audio, resulting in similar facial motions even when the audio signals differ. In contrast, our model is capable of synthesizing a wide range of facial expressions in response to different audio inputs.

Lip, Expression and Head Pose Disentanglement. In 3DMM, the desired motion of face i are expressed with parameter set $p_i \equiv \{\beta_i, R_i, t_i\}$, where the head rotation and translation are expressed as $R \in SO(3)$ and $t \in \mathbb{R}^3$. Consequently, there are six coefficients that govern the control of head pose. Additionally, since we leverage the indices of 3DMM coefficients strongly related to lip movement, we can manipulate either lip movement or facial expression by replacing the corresponding indices with those in driven frames. In a word, our model also enables free control w.r.t pose (show in Figure 1), expression (show in Figure 2) and lip movement with reference videos.

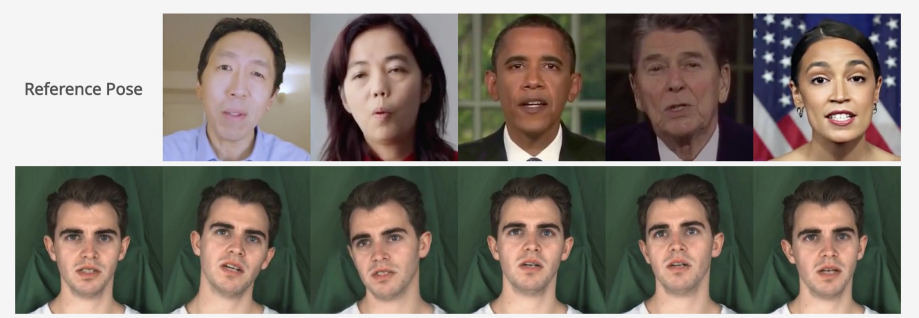

Fig. 1: Control head pose with reference videos.

3 User Study

We conducted a human evaluation on Amazon Mechanical Turk (AMT) to evaluate the qualitative result among SoTA approaches since automatic evaluation metrics cannot adequately measure the naturalness and subtle expressions of generated videos. For each assignment, we present talking head videos which generated by different models with the same audio and neutral identity image as input. The videos were displayed in parallel, with their order randomly shuffled. Each assignment was assessed by three

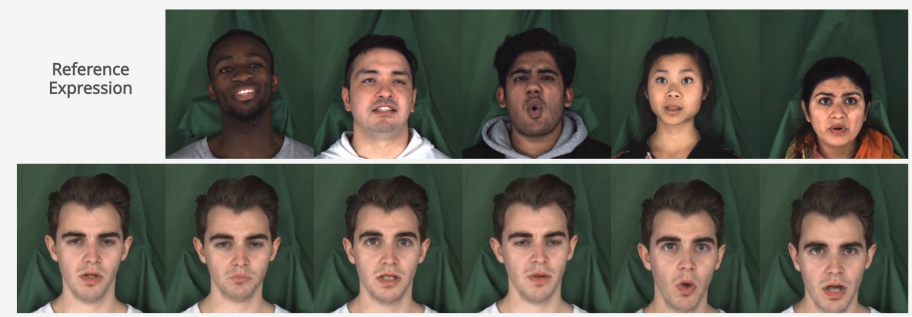

Fig. 2: Control facial expression with reference videos.

Fig. 3: Generated 3DMM with different identity texture.

distinct Mturkers. Participants were asked to select their preferred video based on emotion alignment, lip synchronization, motion diversity, video sharpness, and overall naturalness. Figure 6 shows our designed interface for comparative evaluation between our model and runner-ups.

4 Analysis

Emotion of different intensities. We further investigate the mouth landmark distance (F-LMD) and face landmark distance (M-LMD) for different emotions and intensities and report the results in Table 3. Our findings reveal that the distances associated with sadness and fear are relatively closer to the ground truth, whereas the distance for contempt is significantly higher. We assume that the facial movements associated with contempt emotion exhibit greater flexibility and diversity. In addition, we observed cases where the distance at level 1 exceeds that at level 3, which seems counterproductive. One possible explanation for this discrepancy is that the emotional cues conveyed through audio at level 1 may not be sufficiently pronounced for the model to capture accurately. Conversely, audio at level 3 contains a more intense emotional signal, enabling the model to generate more vivid facial expressions.

Visualize centroid representation We obtain the centroid code embedding in the codebook and pass through the trained VQ-VAE decoder to reconstruct the 3DMM feature. As each code embedding corresponds to a set of four frames, we proceed to randomly select one and generate a visualization of the resulting 3DMM feature, which is visualized in Figure 4.

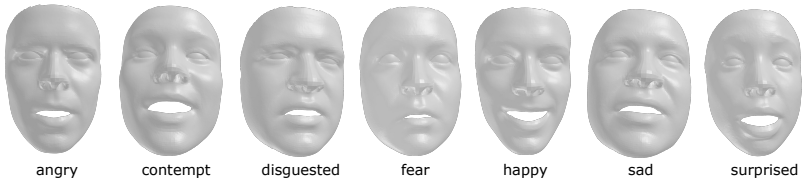

Fig. 4: Reconstructed centroid codes of emotional clustered embedding in the learned codebook.

Raise head phenomenon in MEAD dataset We randomly sample frames from MEAD to show the raise head phenomenon in the original dataset, as shown in Figure 5. (1) In instances where the emotion label of a video is 'contempt', individuals tend to exhibit a tendency to raise their heads to express this emotion. (2) The MEAD dataset comprises actors who read provided transcripts and portray the associated emotions, which leads them to raise their heads and direct their gaze in accordance with the content of the transcript.

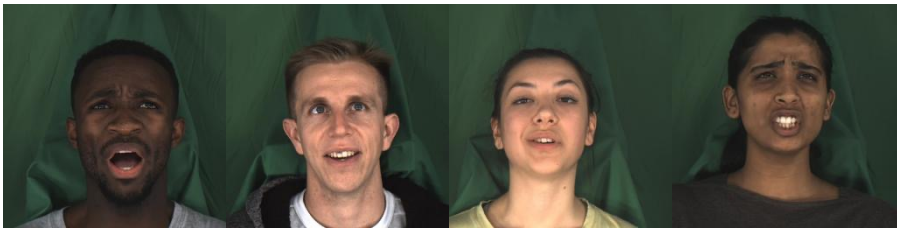

Fig. 5: Raise head phenomenon in MEAD [16] dataset.

5 Talking Head Literature Comparison

Table 2 presents a comprehensive comparison of various audio-driven talking video generation methods. The table showcases the following information: (1) **Type**: the modeling space of each method, including landmark, NeRF, motion Fields, 3DMM and raw image; (2) **Driven Frames**: Whether the method depends on driven frames and how it utilizes them to control specific features in the generated videos, including pose, expression, and style; (3) **Arbitrary Identity**: Whether the model could generate videos with arbitrary identity without any fine-tuning; (4) **Emotion**: Whether the approach takes emotion modeling into consideration and which type of input they need to control emotion, including discrete label or driven frames. Note that our method involves modeling emotions by learning the correlation between audio signals and facial expressions. This allows our model to generate a wide range of natural facial motions aligned with the audio fluctuations, without requiring external labels as input.

6 Implementation Details

Quantization strategy. The VQ-VAE's naive training suffers from codebook collapse, as noted in previous work [10, 15]. To improve codebook utilization, we employ two quantization strategies from [10]: exponential moving average (EMA) and codebook reset (Code Reset). The EMA strategy enables the codebook $\mathcal{C}$ to evolve smoothly over time with the formula $\mathcal{C}_t \leftarrow \lambda\mathcal{C}_{t-1} + (1-\lambda)\mathcal{C}_{t-1}$, where $\mathcal{C}_t$ represents the codebook at iteration t and λ is the exponential moving constant. On the other hand, the Code Reset strategy identifies inactive codes and reassigns them during training.
VQ-VAE. The temporal downsampling and upsampling rate l is 4 of VQ-VAE autoencoder, and the size K of learnable codebook is 1024 of dimension 256. We train the VQ-VAE with AdamW [7] optimizer with warm up learning rate, and with $[\beta_1, \beta_2] = [0.9, 0.99]$, exponential moving constant $\lambda = 0.99$ and batch size 512.
Cross-modal Transformer. The audio feature is extracted by HuBERT model [3] which finetunes on downstream emotion recognition task [21][2]. The context window size h of self-attention temporal bias mask is 4, and the temporal alignment ratio k of cross-attention alignment mask is set to 8. The second stage training is optimized with AdamW [7] with a learning rate $1e-4$ and batch size of 128.
Lip branch. The model is trained with continuous 5 frames and the audio feature for each frame is a 0.2s mel-spectrogram.

References

1. Das, D., Biswas, S., Sinha, S., Bhowmick, B.: Speech-driven facial animation using cascaded gans for learning of motion and texture. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16. pp. 408–424. Springer (2020) 7
2. Guo, Y., Chen, K., Liang, S., Liu, Y.J., Bao, H., Zhang, J.: Ad-nerf: Audio driven neural radiance fields for talking head synthesis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5784–5794 (2021) 7
3. Hsu, W.N., Bolte, B., Tsai, Y.H.H., Lakhotia, K., Salakhutdinov, R., Mohamed, A.: Hubert: Self-supervised speech representation learning by masked prediction of hidden units. IEEE/ACM Transactions on Audio, Speech, and Language Processing **29**, 3451–3460 (2021) 5
4. Ji, X., Zhou, H., Wang, K., Wu, Q., Wu, W., Xu, F., Cao, X.: Eamm: One-shot emotional talking face via audio-based emotion-aware motion model. In: ACM SIGGRAPH 2022 Conference Proceedings. pp. 1–10 (2022) 1, 7
5. Ji, X., Zhou, H., Wang, K., Wu, W., Loy, C.C., Cao, X., Xu, F.: Audio-driven emotional video portraits. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 14080–14089 (2021) 7
6. Liu, X., Xu, Y., Wu, Q., Zhou, H., Wu, W., Zhou, B.: Semantic-aware implicit neural audio-driven video portrait generation. In: Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXVII. pp. 106–125. Springer (2022) 7
7. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017) 5

[2] https://huggingface.co/superb/hubert-base-superb-er

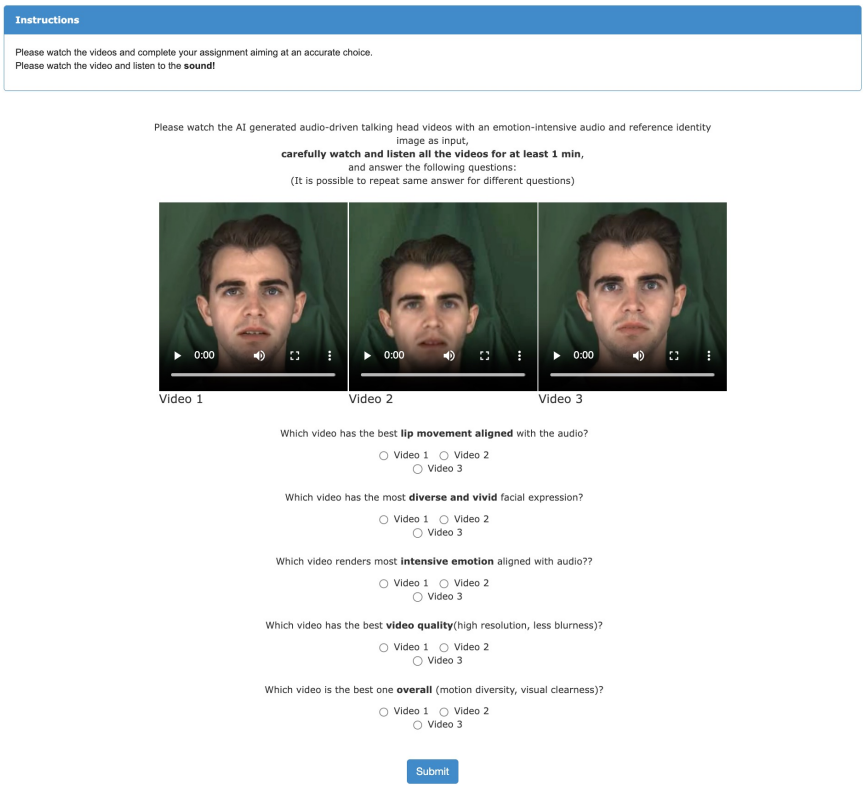

Fig. 6: User study interface on Amazon Mechanical Turk.

8. Ma, Y., Wang, S., Hu, Z., Fan, C., Lv, T., Ding, Y., Deng, Z., Yu, X.: Styletalk: One-shot talking head generation with controllable speaking styles. arXiv preprint arXiv:2301.01081 (2023) 7
9. Prajwal, K., Mukhopadhyay, R., Namboodiri, V.P., Jawahar, C.: 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. pp. 484–492 (2020) 7
10. Razavi, A., Van den Oord, A., Vinyals, O.: Generating diverse high-fidelity images with vq-vae-2. Advances in neural information processing systems **32** (2019) 5
11. Shen, S., Zhao, W., Meng, Z., Li, W., Zhu, Z., Zhou, J., Lu, J.: Difftalk: Crafting diffusion models for generalized talking head synthesis. arXiv preprint arXiv:2301.03786 (2023) 7
12. Sinha, S., Biswas, S., Yadav, R., Bhowmick, B.: Emotion-controllable generalized talking face generation. arXiv preprint arXiv:2205.01155 (2022) 7
13. Song, L., Wu, W., Qian, C., He, R., Loy, C.C.: Everybody's talkin': Let me talk as you want. IEEE Transactions on Information Forensics and Security **17**, 585–598 (2022) 7
14. Stypułkowski, M., Vougioukas, K., He, S., Zięba, M., Petridis, S., Pantic, M.: Diffused heads: Diffusion models beat gans on talking-face generation. arXiv preprint arXiv:2301.03396 (2023) 7
15. Van Den Oord, A., Vinyals, O., et al.: Neural discrete representation learning. Advances in neural information processing systems **30** (2017) 5

Methods	Type	Driven Frames	Arbitrary Identity	Emotion
Das et al. [1]	Landmark	No	✓	–
Wav2Lip [9]	Raw Image	No	✓	–
MakeItTalk [26]	Landmark	No	✓	–
MEAD [16]	Landmark	No	✗	Label
AD-NeRF [2]	NeRF	No	✗	–
HDTF [24]	3DMM	No	✓	–
Audio2head [17]	Motion Fields	No	✓	–
PC-AVS [25]	Raw Image	Pose	✓	–
Song et al. [13]	3DMM	Expression	✓	–
EVP [5]	3DMM	No	✓	Label
Wu et al. [20]	3DMM	Style	✓	–
AVCT [18]	Motion Fields	No	✓	–
SSP-NeRF [6]	NeRF	No	✗	–
Sinha et al. [12]	Landmark	No	✓	Label
EAMM [4]	Motion Fields	Expression	✓	Frames
StyleTalk [8]	3DMM	Style	✓	–
Diffused Heads [14]	Raw Image	No	✓	Label
DiffTalk [11]	Raw Image	Expression	✓	–
GeneFace [22]	Landmark & NeRF	Pose	✓	–
SadTalker [23]	3DMM	No	✓	–
EmoTalker (ours)	3DMM	No	✓	No need input!

Table 2: Comprehensive comparison among audio-driven talking video generation methods. '–' means those methods do not take emotion into consideration. Note that our method does not need extra emotional information (frames or label) as input but implicitly model emotional signal from audio instead.

F-LMD ↓	**Emotion**						
	angry	contempt	disgusted	fear	happy	sad	surprised
Level 1	4.847	5.028	4.038	3.753	4.254	4.277	4.324
Level 2	4.483	5.078	5.123	4.706	4.014	4.317	4.693
Level 3	4.214	5.355	5.391	4.912	4.567	3.246	4.809
M-LMD ↓	angry	contempt	disgusted	fear	happy	sad	surprised
Level 1	6.379	6.479	4.354	3.849	5.110	4.855	4.88
Level 2	5.199	6.154	5.941	5.288	4.484	5.272	5.195
Level 3	4.465	5.972	6.104	5.774	5.219	3.604	5.282

Table 3: Landmark distance of our method among different emotions and intensities.

16. Wang, K., Wu, Q., Song, L., Yang, Z., Wu, W., Qian, C., He, R., Qiao, Y., Loy, C.C.: 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. pp. 700–717. Springer (2020) 4, 7
17. Wang, S., Li, L., Ding, Y., Fan, C., Yu, X.: Audio2head: Audio-driven one-shot talking-head generation with natural head motion. arXiv preprint arXiv:2107.09293 (2021) 7
18. Wang, S., Li, L., Ding, Y., Yu, X.: One-shot talking face generation from single-speaker audio-visual correlation learning. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 36, pp. 2531–2539 (2022) 7
19. Wang, X., Li, Y., Zhang, H., Shan, Y.: Towards real-world blind face restoration with generative facial prior. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9168–9178 (2021) 2
20. Wu, H., Jia, J., Wang, H., Dou, Y., Duan, C., Deng, Q.: Imitating arbitrary talking style for realistic audio-driven talking face synthesis. In: Proceedings of the 29th ACM International Conference on Multimedia. pp. 1478–1486 (2021) 7
21. Yang, S.w., Chi, P.H., Chuang, Y.S., Lai, C.I.J., Lakhotia, K., Lin, Y.Y., Liu, A.T., Shi, J., Chang, X., Lin, G.T., et al.: Superb: Speech processing universal performance benchmark. arXiv preprint arXiv:2105.01051 (2021) 5
22. Ye, Z., Jiang, Z., Ren, Y., Liu, J., He, J., Zhao, Z.: Geneface: Generalized and high-fidelity audio-driven 3d talking face synthesis. arXiv preprint arXiv:2301.13430 (2023) 7
23. Zhang, W., Cun, X., Wang, X., Zhang, Y., Shen, X., Guo, Y., Shan, Y., Wang, F.: Sadtalker: Learning realistic 3d motion coefficients for stylized audio-driven single image talking face animation. arXiv preprint arXiv:2211.12194 (2022) 7
24. Zhang, Z., Li, L., Ding, Y., Fan, C.: 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. pp. 3661–3670 (2021) 7
25. Zhou, H., Sun, Y., Wu, W., Loy, C.C., Wang, X., Liu, Z.: Pose-controllable talking face generation by implicitly modularized audio-visual representation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4176–4186 (2021) 7
26. Zhou, Y., Han, X., Shechtman, E., Echevarria, J., Kalogerakis, E., Li, D.: Makelttalk: speaker-aware talking-head animation. ACM Transactions On Graphics (TOG) **39**(6), 1–15 (2020) 7