

# PC-Talk: Precise Facial Animation Control for Audio-Driven Talking Face Generation

## Supplementary Material

In the supplementary material, we mainly focus on the following aspects:

1. **Further Analysis:** We examine more features and provide further explanations of our framework, including style space visualization, motivation of speaking style editing and a deeper look into the implicit keypoint representation and two-stage generation process.
2. **Additional Experiments:** We present comparisons on additional datasets and conduct an ablation study on components within the LAC module.
3. **Implement Details:** We provide detailed information about our framework, covering model architecture, emotion sources, loss functions, and training details.
4. **Discussion:** We discuss potential improvements for future work and address ethical considerations.

Additionally, we strongly encourage watching the supplementary video, which showcases comparisons with other methods and demonstrates the controllability of speaking style and emotional expression. Our method clearly outperforms others in overall realism and control precision.

### 1. Further Analysis

**Style Space Visualize.** As shown in Fig. S1(b), we project style embedding video clips and one-hot codes into the style space and visualize the results using PCA. We observe that embeddings corresponding to the same speaking style form tight clusters, while remaining well separated across different identities (since each identity exhibits a distinct speaking style). This demonstrates the consistency and effectiveness of our learned style space, which is derived from both video references and preset one-hot inputs.

**Speaking Style Editing.** The motivation behind speaking style editing is to explicitly define and manipulate speaking styles. When the desired style cannot be achieved through either a preset option or a video reference, we edit the speaking style using specific lip articulations to simulate particular speaking habits. This approach differs from simply scaling lip movements, as it selectively targets specific articulation. However, the extent of style editing is constrained to maintain accurate lip-sync performance.

**Implicit Keypoint.** In our framework, we adopt implicit keypoints as the intermediate representation due to their balanced visual quality and computational efficiency—both crucial for our application. While EAT [6] utilizes implicit keypoints from Face-vid2vid[14], it offers limited fine-grained control, such as in eyebrow movement. In contrast, our approach employs implicit keypoints similar to

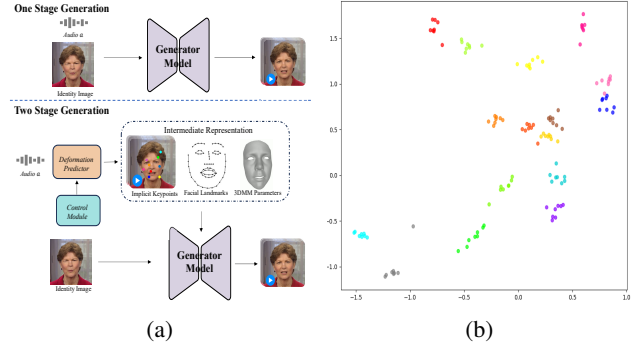


Figure S1. (a) Two stage generation. (b) Style Space Visualize. Each color represents a speaking style sample from both one-hot code and reference video.

those in LivePortrait[7], a video-driven method, which connect several implicit keypoints to 2D facial landmarks, enabling independent animation of each facial region. Specifically, we introduce a landmark-guided loss during training, which optimizes the projected distance between implicit keypoints and 2D facial landmarks. Our approach generates vivid talking faces with expressive emotions from audio alone, enabling precise editing of speaking style and emotional expression—capabilities notably absent in existing methods.

**Two-stage Generation.** As depicted in Fig. S1(a), our method is not theoretically limited to a single implicit keypoint framework, but can be extended to various two-stage generation approaches, including advanced diffusion models that offer greater generative capability at the expense of efficiency. The control module can be trained not only on implicit keypoints, but also on other intermediate representations such as facial landmarks or 3DMM parameters. Once the desired speaking style and emotional expression representations are obtained from the control module, they are passed to the generation model for final output.

Our model supports not only realistic human faces but also styles like cartoons and paintings. This versatility stems from training the implicit keypoints framework with a combined dataset of human and stylized faces. It can even be extended to animals, such as cats and dogs, showing the strong generalize ability.

### 2. Additional Experiments

#### 2.1. Experiments on other Datasets

As shown in Tab. S1, we randomly selected 50 video clips from the VoxCeleb2[3] and LRW[15] datasets to com-

Table S1. Experiments on VoxCeleb2 and LRW dataset

Method	VoxCeleb2					LRW				
	LSE-C $\uparrow$	LSE-D $\downarrow$	FID $\downarrow$	NIQE $\downarrow$	FVD $\downarrow$	LSE-C $\uparrow$	LSE-D $\downarrow$	FID $\downarrow$	NIQE $\downarrow$	FVD $\downarrow$
Wav2Lip[10]	<u>7.65</u>	<b>7.13</b>	42.16	47.18	275.83	<u>6.32</u>	<u>8.31</u>	39.24	56.22	176.09
MuseTalk[17]	4.07	9.74	<u>41.21</u>	47.46	<u>235.26</u>	3.73	10.47	<u>25.23</u>	55.48	<b>153.81</b>
SadTalker[16]	6.08	8.03	64.64	34.69	455.95	5.35	8.71	48.58	15.39	516.84
Hallo-v2[4]	7.26	8.17	80.34	<b>13.29</b>	359.60	6.28	9.13	57.03	<u>13.83</u>	483.89
<b>PC-Talk(Ours)</b>	<b>7.74</b>	<u>7.52</u>	<b>32.37</b>	<u>13.34</u>	<b>205.55</b>	<b>7.14</b>	<b>8.04</b>	<b>24.83</b>	<b>12.48</b>	<u>163.70</u>

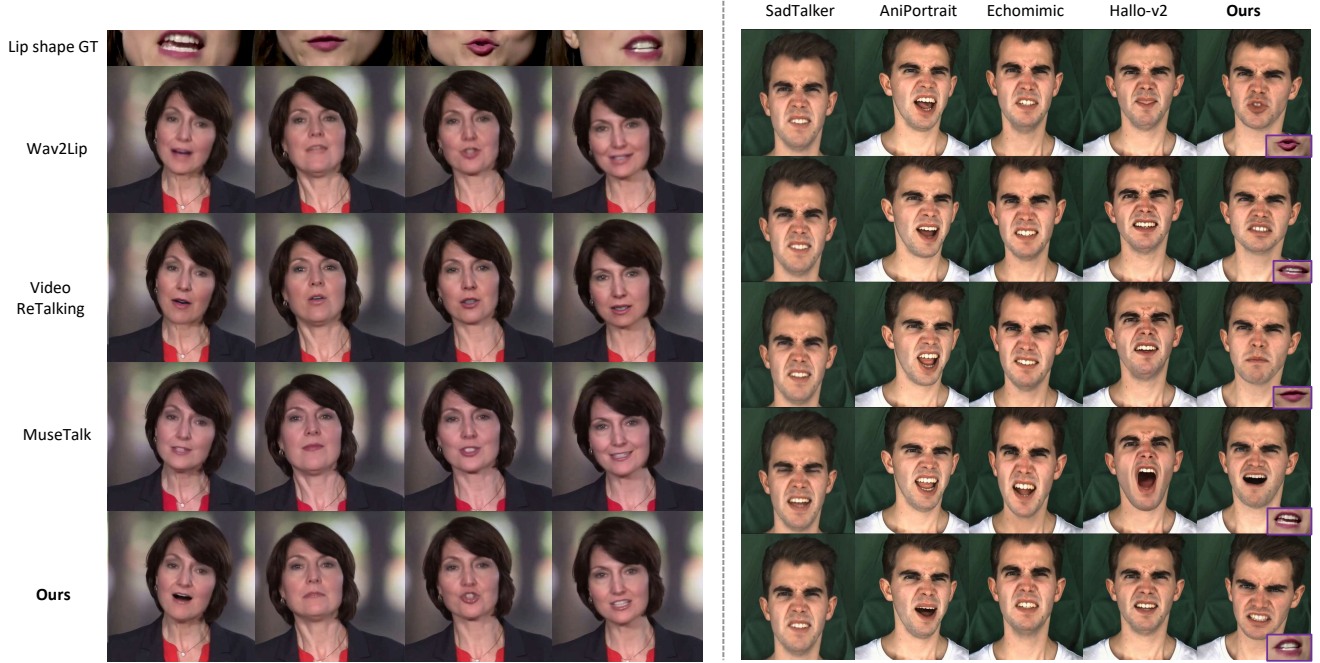


Figure S2. Comparison with other baseline. The lip shape GT in right side is at the corner of each row.

hensively evaluate the performance of our method. These datasets feature diverse speakers, varying head poses, lighting conditions, and background complexity, making them representative of challenging in-the-wild scenarios. Our approach consistently achieved the highest rank across nearly all evaluation metrics, demonstrating not only superior overall performance but also strong generalization ability. These results highlight the robustness and adaptability of our method in real-world, unconstrained environments.

## 2.2. More Ablation Study

We conducted a series of ablation studies focusing on key components related to the lip-audio alignment module, including the use of lip-specific keypoints, the incorporation of a final MLP layer, and the effect of data augmentation strategies. As shown in Tab. S2, we observe that restricting prediction to only lip-related keypoints leads to improved lip-synchronization accuracy, due to reduced ambiguity and more focused modeling of lip motion. Incorporating the final MLP layer into our architecture significantly enhances

synchronization quality, as it is trained on implicit keypoints rather than solely on expression deformations. Furthermore, applying data augmentation contributes to improved generalization, further boosting overall performance.

## 2.3. More Visible Results

As shown in Fig. S2, our methods demonstrate superior results compared to other approaches. Both lip synchronization and image quality are significantly improved. Additionally, as shown in Fig. S3, the emotional expressiveness of our results is distinct. Our methods effectively support various facial styles in Fig. S6, including not only realistic human faces but also paintings and cartoons, showcasing their versatility and effectiveness.

## 3. Implemtation Details

### 3.1. Model Architecture

As shown in Fig. S4, our model adopts an auto-regressive method adapted from FaceFormer [5] for expression predic-



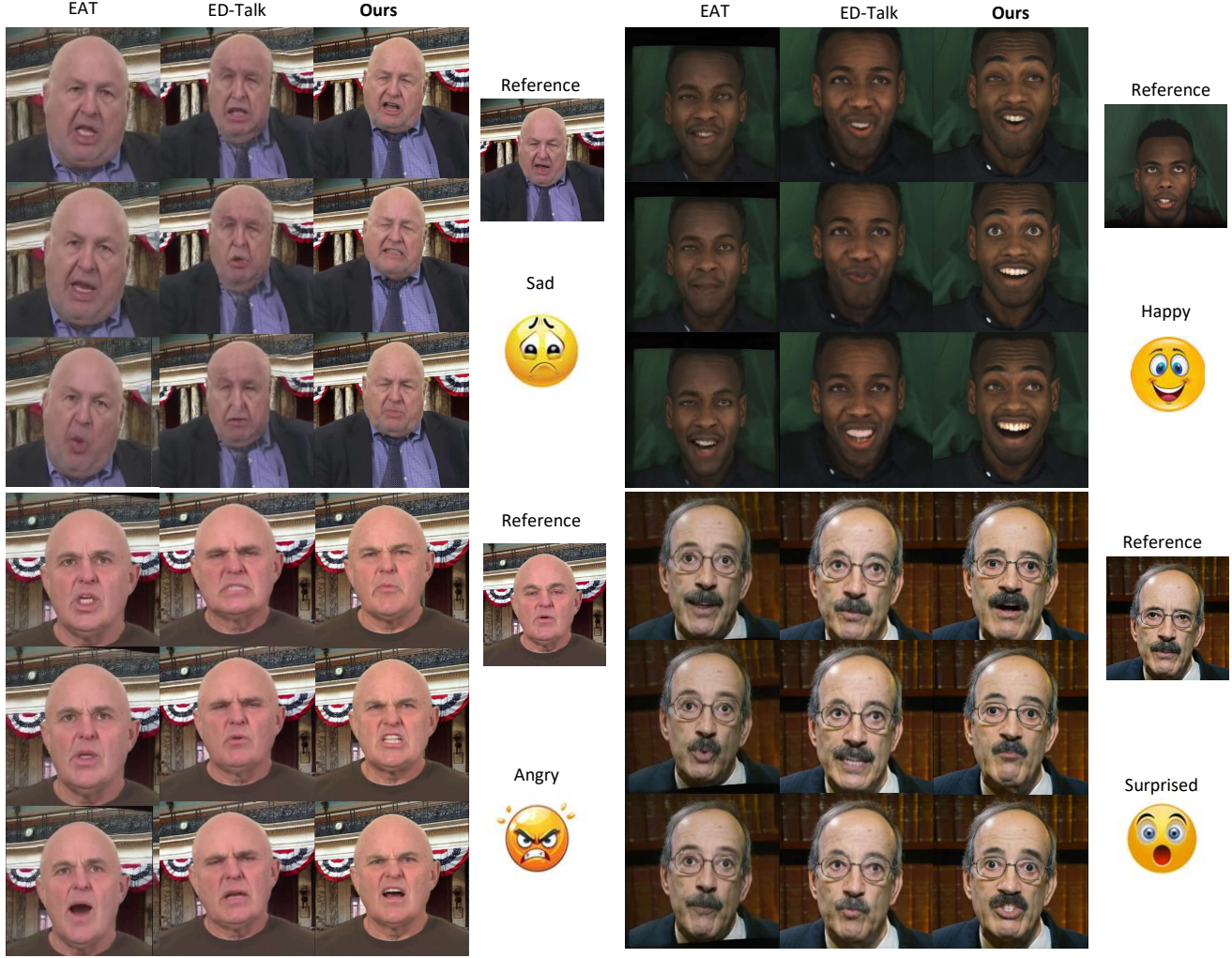


Figure S3. Comparison with other emotional talking face generation methods.

tion. The speaking style can be specified in two ways: either from a reference video or via a preset option provided in the dataset. To extract style information, we use a Transformer Encoder that encodes a sequence of expression deformations—obtained from the reference video—into a style embedding. Preset styles are represented as one-hot vectors, which are projected into the same style space as the former. During inference, users can upload a reference video to enable speaking style adaptation, or switch to the preset mode when no video is available.

The overall architecture follows a Transformer Decoder design [13], consisting of a self-attention layer that captures temporal coherence across frames and a cross-attention layer that incorporates the audio embedding  $e_a$ . An MLP layer is applied to refine the final output. The autoregressive mechanism contributes to the temporal stability of the generated video. This architecture is also integrated into the emotion control module, serving as a unified predic-

tor for generating expression deformations that reflect combined emotional deformation.

### 3.2. Various Emotion Source

Our method leverages multiple sources to flexibly control the emotional expressions of a talking face. For direct control sources like images, we replace the original facial expression with the one extracted from the source image, facilitating straightforward and effective emotion transfer. Additionally, we utilize semantic-aware implicit keypoints to transfer expressions to specific facial regions selectively.

For more complex control sources, such as audio or text, we derive emotional embeddings from these sources. We employ pre-trained emotion recognition models to extract features and map them onto our generation network as emotional embeddings. For example, we use the Wav2Vec [11] model for audio sources and Emoberta [8] for text sources to inform our emotional generation. For video references, we

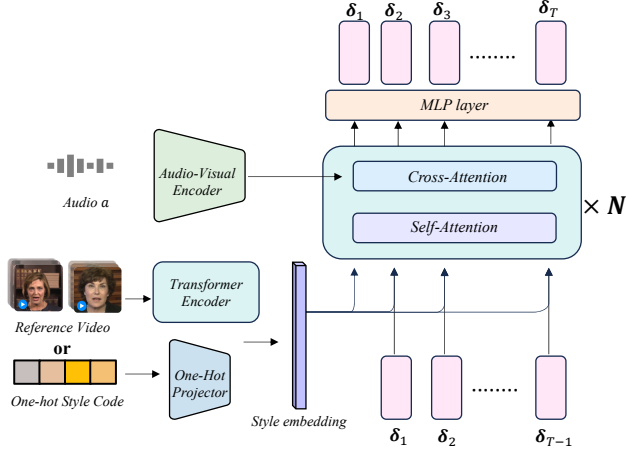


Figure S4. Architecture of our style-aware expression predictor. We use an auto-regressive model to predict expression parameters  $\delta_{pred}$  from audio input  $a$  and speaking style source  $s$ , supporting both video reference and preset one-hot style code as input.

simply use an unpretrained Transformer [13] with expression input and can incorporate the average of emotion deformation from the video reference as direct control. Once these emotional embeddings are obtained, they are integrated into the generation process to produce an emotionally expressive talking face.

### 3.3. Loss Function

We incorporate various types of loss functions when training the Lip-Audio Alignment Control (LAC) module. The style-aware expression predictor generates a predicted keypoint set  $K_{pred}$ , which is subtracted from the original keypoint set  $K_{ori}$  to produce the final lip-sync deformation  $D_l$ . The overall loss function  $\mathcal{L}_{LAC}$  is formulated as:

$$\mathcal{L}_{LAC} = \mathcal{L}_{sync} + \lambda_{kp} \mathcal{L}_{kp} + \lambda_{reg} \mathcal{L}_{reg} + \lambda_{vel} \mathcal{L}_{vel} + \lambda_{style} \mathcal{L}_{style}, \quad (1)$$

The sync loss  $\mathcal{L}_{sync}$  is adapted from Wav2Lip [10], which significantly enhances the model’s ability to achieve accurate lip synchronization. The  $\mathcal{L}_{sync}$  is formulated as:

$$\mathcal{L}_{sync} = -\frac{\mathbf{S}_v(I_{gt:gt+4})^T \cdot \mathbf{S}_a(a_{gt:gt+4})}{\|\mathbf{S}_v(I_{gt:gt+4})\|_2 \|\mathbf{S}_a(a_{gt:gt+4})\|_2}, \quad (2)$$

where  $I_{gt:gt+4}$  is a sequence of frames as image input, and  $a_{gt:gt+4}$  is the audio input. The  $\mathcal{L}_{kp}$  and  $\mathcal{L}_{reg}$  are employed to constrain excessive deformation changes, ensuring stable and natural results, which is formulated as:

$$\mathcal{L}_{kp} = \sum_{t=t_0}^{t_0+4} \|K_{pred}^t - K_{gt}^t\|_2, \quad (3)$$

$$\mathcal{L}_{reg} = \sum_{t=t_0}^{t_0+4} \|K_{pred}^t\|_2, \quad (4)$$

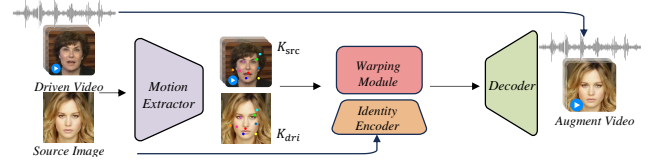


Figure S5. Data augmentation using video-driven portrait animation from same framework.

Table S2. Ablation Study on lip-audio alignment.

Lip-related Keypoint	✗	✓	✓	✓
Final MLP Layer	✗	✗	✓	✓
Data Augmentation	✗	✗	✗	✓
LSE-C (↑)	8.26	8.47	9.26	<b>9.37</b>

$\mathcal{L}_{vel}$  is employed to ensure temporal consistency, which is formulated as:

$$\mathcal{L}_{vel} = \sum_{t=t_0+1}^{t_1} \left\| (\delta_{pred}^t - \delta_{pred}^{t-1}) - (\delta_{gt}^t - \delta_{gt}^{t-1}) \right\|_2, \quad (5)$$

Additionally, we incorporate a style discrimination loss  $\mathcal{L}_{style}$  to ensure that the model adapts the style from the source effectively. A pretrained classifier is used to supervise the generated results. The  $\mathcal{L}_{style}$  is formulated as:

$$L_{style} = -\log p_i, \quad (6)$$

where  $i$  is category of the speaking style. Note that we only calculate with ground truth category, as their are similar speaking style in dataset which is hard to identify.

For emotion control module, the loss function is identical with  $\mathcal{L}_{kp}$  in LAC module. Additionally, we add a classifier loss  $\mathcal{L}_{cls}$  on emotion embedding to generate emotional expressive feature from various emotion source.

$$L_{cls} = -\sum_{c=1}^M (\mathbf{y}_c * \log \mathbf{p}_c), \quad (7)$$

where  $M$  is numbers of emotion categories,  $\mathbf{y}_c$  is one-hot embedding carries the emotion label  $c$ , and  $\mathbf{p}_c$  denotes the predicted probability that belongs to class  $c$ .

### 3.4. Training Details

We preprocess the dataset by converting videos to 25 fps and sampling audio at 16 kHz. We initialize frozen parameters including motion extractor, identity encoder, warping module, and decoder from LivePortrait [7], while the audio encoder is adopted from Wav2Lip [10]. Implicit keypoints and audio embeddings are extracted using these components before training. The LAC module and EMC module are trained separately using the Adam optimizer [9] with a learning rate of  $1e-4$ .



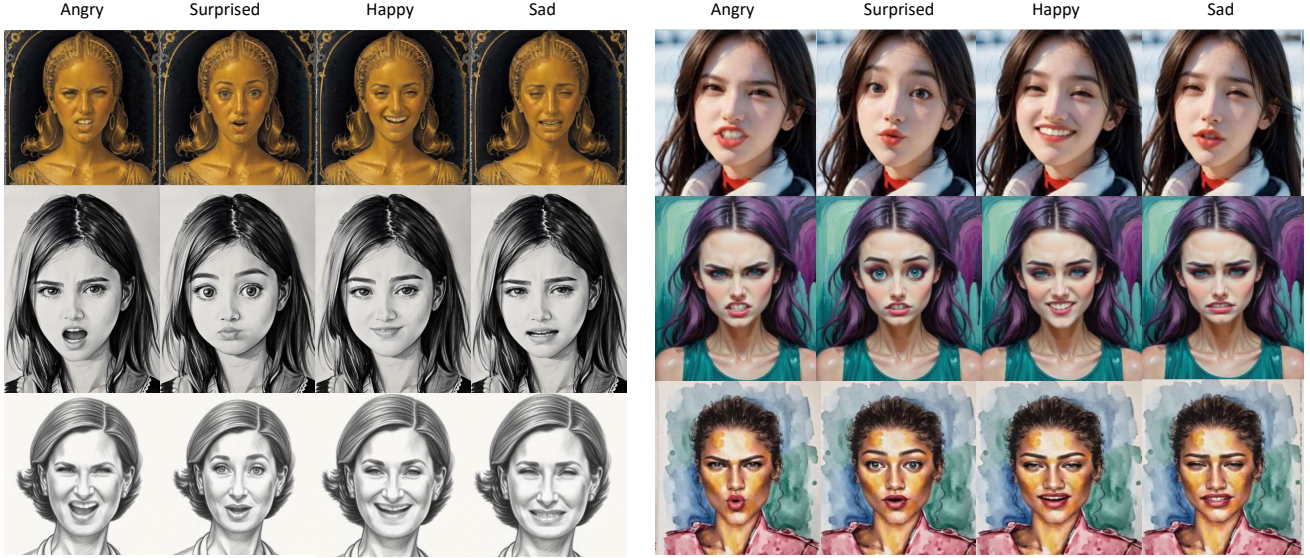


Figure S6. Comparison with other emotional talking face generation methods.

For the LAC module, the style-aware expression predictor is trained over 24 hours with a window size of 50, focusing on implicit keypoints. Then we transition to the additional MLP layer, using the comprehensive network training which is conducted within the image space to compute  $\mathcal{L}_{sync}$ , necessitating a reduction in window size to 5 and extending the training time to 72 hours. As shown in Fig. S5, we employed a data augmentation through unsupervised video-driven portrait animation to enhance the training effectiveness. We use HDTF [18] as the driving video coupled with audio tracks. We filter this with a SyncNet [2] to ensure the lip-sync quality of data augmentation. This unsupervised augmented data ensures a fair comparison across experiments.

Given that the EMC module might adversely affect lip synchronization, we refine the lip synchronization using the lip refiner from LAC module to improve performance while maintaining emotional expression. A Kalman filter is used on implicit keypoints deformation during inference to degrade instability. Our framework achieves an impressive 30 frames per second, thus demonstrating high efficiency and enabling real-time generation.

## 4. Discussion

**Pose Generation:** Current methods extract pose information directly from video or through pre-defined templates to apply to the input image, ensuring pose uniformity. These methods utilize relative pose when working with images. For future work, we propose generating pose information directly from a noise vector  $z$ . Additionally, incorporating audio as an input may prove beneficial, as the tone of voice can reflect head movement.

**Disentanglement of Implicit Keypoints:** The quality of facial representation is critical in our methods. Although the current implicit keypoints already set a high standard, there is room for improvement, especially in handling large poses and cross-identity inference. The disentanglement of identity, pose, and appearance features is crucial for enhancing quality. To achieve more accurate implicit keypoints, integrating 3D supervision might provide better results in future iterations.

**Ethical Consideration.** Our method can generate high-fidelity talking faces with fine-grained control over speaking style and emotion. We acknowledge the potential risks associated with the misuse of this technology on public platforms. In light of these concerns, we are committed to responsibly sharing our results to support the development of deepfake detection methods and promote safe, ethical use.

## References

- [1] Zhiyuan Chen, Jiajiong Cao, Zhiqian Chen, Yuming Li, and Chenguang Ma. Echomimic: Lifelike audio-driven portrait animations through editable landmark conditions. *arXiv preprint arXiv:2407.08136*, 2024.
- [2] J. S. Chung and A. Zisserman. Out of time: automated lip sync in the wild. In *Workshop on Multi-view Lip-reading, ACCV*, 2016. 6
- [3] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. Voxceleb2: Deep speaker recognition. *arXiv preprint arXiv:1806.05622*, 2018. 2
- [4] Jiahao Cui, Hui Li, Yao Yao, Hao Zhu, Hanlin Shang, Kaihui Cheng, Hang Zhou, Siyu Zhu, and Jingdong Wang. Hallo2: Long-duration and high-resolution audio-driven portrait image animation. *arXiv preprint arXiv:2410.07718*, 2024. 3
- [5] Yingruo Fan, Zhaojiang Lin, Jun Saito, Wenping Wang, and

- Taku Komura. Faceformer: Speech-driven 3d facial animation with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18770–18780, 2022. 3
- [6] 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
- [7] Jianzhu Guo, Dingyun Zhang, Xiaoqiang Liu, Zhizhou Zhong, Yuan Zhang, Pengfei Wan, and Di Zhang. Liveportrait: Efficient portrait animation with stitching and retargeting control. *arXiv preprint arXiv:2407.03168*, 2024. 2, 5
- [8] Taewoon Kim and Piek Vossen. Emoberta: Speaker-aware emotion recognition in conversation with roberta. *arXiv preprint arXiv:2108.12009*, 2021. 4
- [9] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5
- [10] KR Prajwal, Rudrabha Mukhopadhyay, Vinay P Nambodiri, 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. 3, 5
- [11] Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. wav2vec: Unsupervised pre-training for speech recognition. *arXiv preprint arXiv:1904.05862*, 2019. 4
- [12] Shuai Tan, Bin Ji, Mengxiao Bi, and Ye Pan. Edtalk: Efficient disentanglement for emotional talking head synthesis. In *European Conference on Computer Vision*, pages 398–416. Springer, 2024.
- [13] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 4, 5
- [14] 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. 2
- [15] Shuang Yang, Yuanhang Zhang, Dalu Feng, Mingmin Yang, Chenhao Wang, Jingyun Xiao, Keyu Long, Shiguang Shan, and Xilin Chen. Lrw-1000: A naturally-distributed large-scale benchmark for lip reading in the wild. In *2019 14th IEEE international conference on automatic face & gesture recognition (FG 2019)*, pages 1–8. IEEE, 2019. 2
- [16] 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. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8652–8661, 2023. 3
- [17] Yue Zhang, Minhao Liu, Zhaokang Chen, Bin Wu, Yubin Zeng, Chao Zhan, Yingjie He, Junxin Huang, and Wenjiang Zhou. Musetalk: Real-time high quality lip synchronization with latent space inpainting. *arXiv preprint arXiv:2410.10122*, 2024. 3
- [18] 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. 6