

In this Appendix, we first provide the user study conducted on MEAD and DH-FaceEmoVid-150 (Appendix D). Then we detail the Mixture of Emotion Experts module and the Emotion-to-Latents module (Appendix A). Next, we introduce the training and testing details (Appendix B). Finally, we provide more visualization results (Appendix C) and further visualizations of the dataset (Appendix E).

A. The Networks Details

Mixture of Emotion Experts In our model, we employ the Mixture of Emotion Experts (MoEE) framework based on cross-attention. Different emotion experts are guided by emotional signals extracted from audio and text to generate single basic emotions. The emotion latent obtained from the Emotion-to-Latents module serves as the key and value in the cross-attention, with a dimension of $(bs, t_{emotion}, c_{emotion})$. Here, bs represents the batch size, $t_{emotion}$ denotes the number of tokens (set to 8), and $c_{emotion}$ refers to the channel dimension of the emotion latent (set to 512). The hidden state is used as the query, which is normalized using LayerNorm before entering the cross-attention module. Finally, a skip connection is applied to prevent issues such as gradient vanishing and gradient explosion, thereby accelerating the convergence of the model.

Emotion-to-Latents To achieve control signals across multiple modalities, we introduce the Emotion-to-Latents module. Specifically, we first encode text, label, and audio inputs using pretrained encoders. To map signals from different modalities into a unified emotion latent space, we incorporate a cross-attention module. Four separate fully connected networks are trained to project the channel dimensions of different embeddings into 512 dimensions, which are used as queries. We also define learnable embeddings that serve as keys and values for attention computation. These embeddings, with a channel dimension of 768, are used to compute a new feature representation based on the learnable embeddings. The resulting emotion latents obtained from the cross-attention module are then fed into the UNet for further processing.

B. Training and Testing Details

Experiments for training and inference were conducted on a platform with 8 NVIDIA A800 GPUs. Each of the training stages consisted of 30,000 steps, utilizing a batch size of 4 and video dimensions of 512×512 pixels. The AdamW [3] optimizer is employed with a learning rate of $1e-5$, and the motion module was initialized using pretrained weights from Animatediff. Each training instance in the second stage produced 14 video frames, with the motion module’s latents concatenated with the first 2 ground truth frames for video continuity. For the diffusion model, a quadratic β schedule is set with $\beta_{min} = 0.05$ and $\beta_{max} = 20$.

Table 1. User Study for MoEE and other baselines on MEAD. The bold values indicate the best results.

Method	Emo.↑	Lip.↑	Nat.↑	ID.↑
AniPortrait [6]	2.05	3.98	4.33	4.31
Hallo [7]	2.03	4.48	4.25	4.05
Echomimic [1]	2.88	4.71	4.65	4.25
StyleTalk [4]	4.44	3.22	3.54	3.58
EAT [2]	4.23	3.57	3.81	4.29
(Ours)	4.65	4.74	4.71	4.55

Table 2. User Study for MoEE and other baselines on DH-FaceEmoVid-150. The bold values indicate the best results.

Method	Emo.↑	Lip.↑	Nat.↑	ID.↑
AniPortrait [6]	2.25	4.02	4.18	4.31
Hallo [7]	2.13	4.21	4.54	4.33
Echomimic [1]	2.78	4.56	4.71	4.25
StyleTalk [4]	4.62	3.51	3.73	3.58
EAT [2]	4.51	3.45	3.75	4.21
(Ours)	4.73	4.81	4.76	4.62

In inference, the model follows the DDIM [5] approach and samples 150 steps. Continuity across sequences are ensured by concatenating noisy latents with feature maps of the last 2 motion frames from the previous step within the motion module.

C. More Visualization Results

To further support the conclusions drawn in the main paper, we provide additional results in this section. Figure 1 shows more video generation results of the proposed approach with different portrait styles. Figure 2 presents additional examples of the single emotion, compound emotion, and AU controls.

D. User Study

We conduct a user study to compare our method with the other methods. We involve 15 experienced users to score the generation quality and controllability of each model. We randomly select 10 generation examples from the MEAD and DH-FaceEmoVid-150 datasets. Our evaluation metric is the Mean Opinion Score (MOS). We assess the lip-sync quality (Lip.), emotion controllability (Emo.), naturalness (Nat.) and identity preservation (ID.). Participants were presented with one video at a time and asked to rate each video for each score on a scale of 1 to 5. We calculated the average score as the final result. As shown in Table 1 and Table 2, Our proposed method achieves the best results across all evaluation criteria.

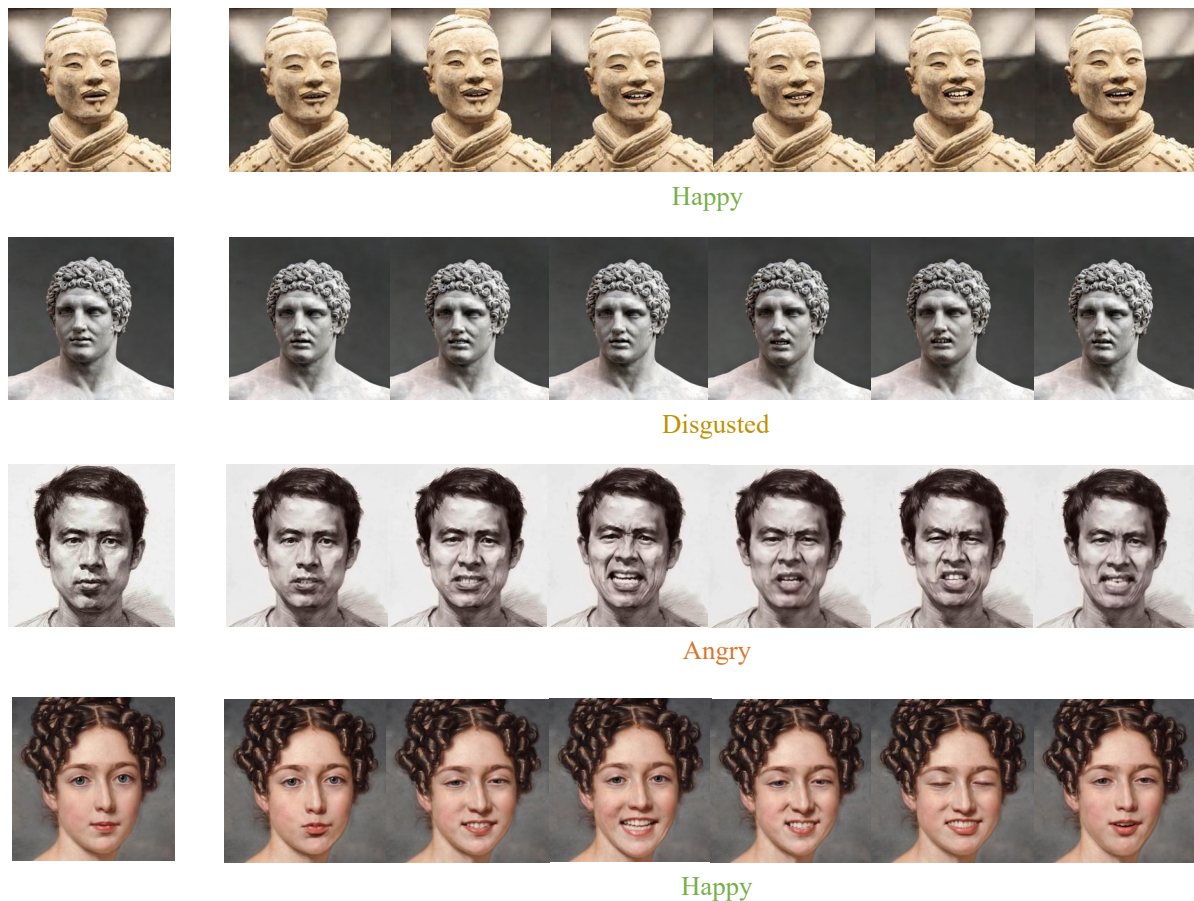


Figure 1. Talking head videos with different portrait styles under various emotions.

E. More Visualization of the Dataset

We present additional samples from the DH-FaceEmoVid-150 dataset, including six basic emotions in Figure 3 and four compound emotions in Figure 4. The FPS of the videos in the dataset has been standardized to 30, with each video clip having a duration of 30 seconds.

References

- [1] Zhiyuan Chen, Jiajiong Cao, Zhiquan Chen, Yuming Li, and Chenguang Ma. Echomimic: Lifelike audio-driven portrait animations through editable landmark conditions. *arXiv preprint arXiv:2407.08136*, 2024. [1](#)
- [2] 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. [1](#)
- [3] I Loshchilov. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. [1](#)
- [4] 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. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1896–1904, 2023. [1](#)
- [5] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. [1](#)
- [6] Huawei Wei, Zejun Yang, and Zhisheng Wang. Aniportrait: Audio-driven synthesis of photorealistic portrait animation. *arXiv preprint arXiv:2403.17694*, 2024. [1](#)
- [7] Mingwang Xu, Hui Li, Qingkun Su, Hanlin Shang, Liwei Zhang, Ce Liu, Jingdong Wang, Yao Yao, and Siyu Zhu. Hallo: Hierarchical audio-driven visual synthesis for portrait image animation. *arXiv preprint arXiv:2406.08801*, 2024. [1](#)

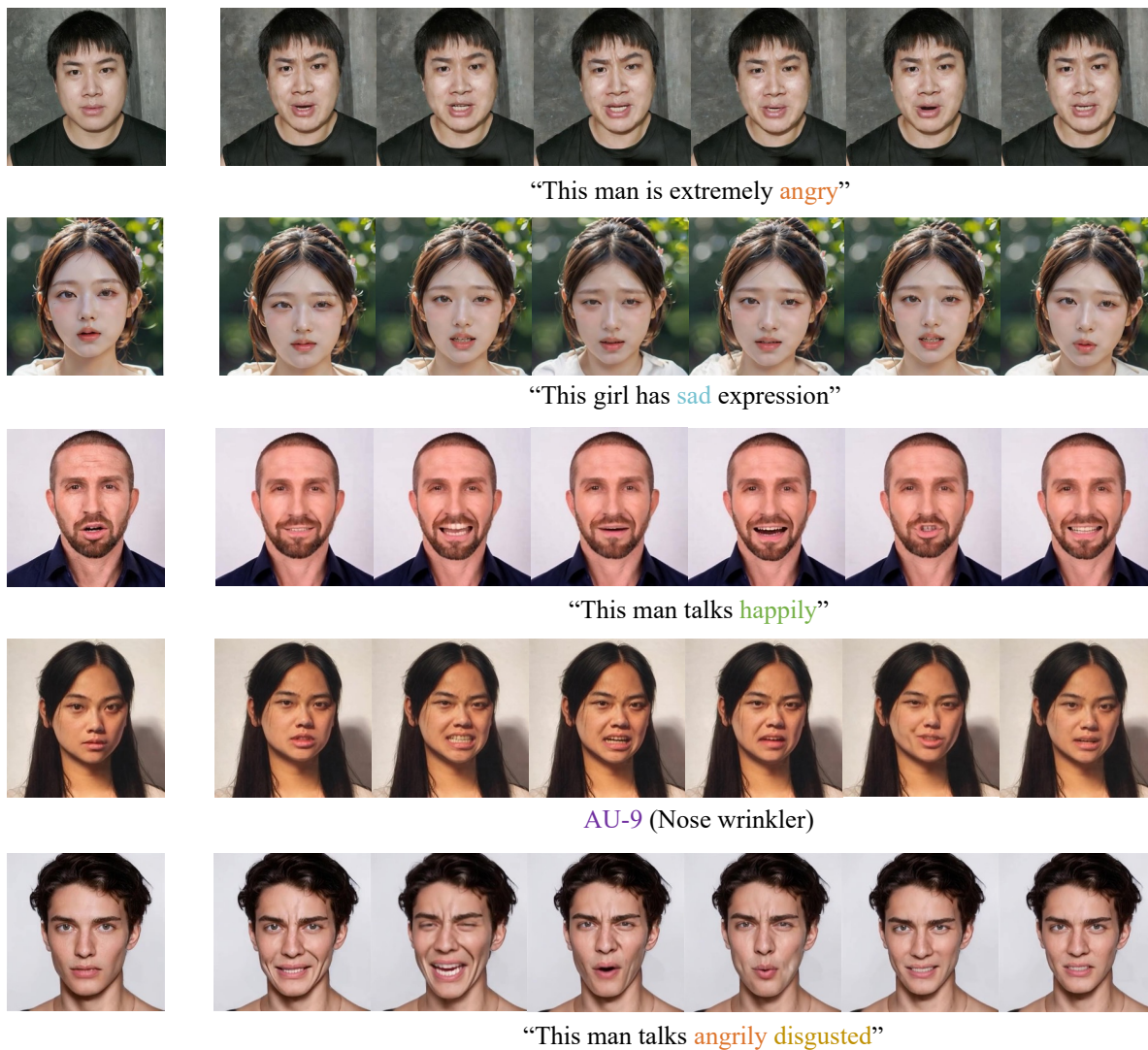


Figure 2. Portrait image animation results on Emotion Control and AU Control.



Figure 3. Visualization of basic emotion examples from the DH-FaceEmoVid-150 dataset.

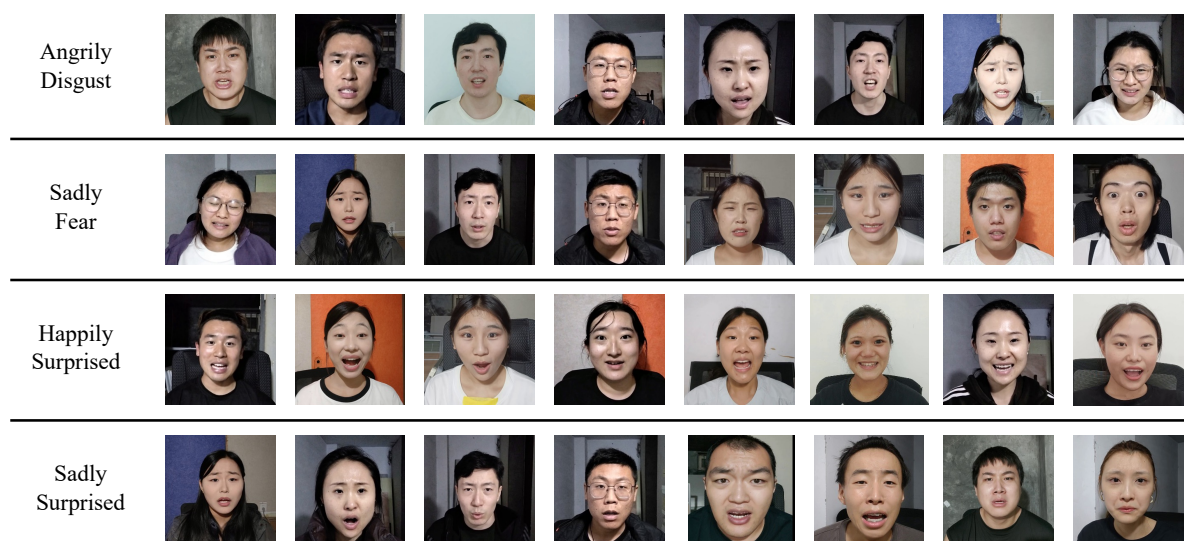


Figure 4. Visualization of compound emotion examples from the DH-FaceEmoVid-150 dataset.