DreamTalk: When Expressive Talking Head Generation Meets Diffusion Probabilistic Models

Yifeng Ma[1]*, Shiwei Zhang[2], Jiayu Wang[2], Xiang Wang[3]*, Yingya Zhang[2], Zhidong Deng[1]

[1] Department of Computer Science and Technology, BNRist, THUAI,
State Key Laboratory of Intelligent Technology and Systems, Tsinghua University
[2]Institute for Intelligent Computing, Alibaba Group
[3]Huazhong University of Science and Technology

mayf18@mails.tsinghua.edu.cn, wxiang@hust.edu.cn, michael@tsinghua.edu.cn
{zhangjin.zsw, wangjiayu.wjy, yingya.zyy}@alibaba-inc.com
https://dreamtalk-project.github.io

Abstract

Diffusion models have shown remarkable success in a variety of downstream generative tasks, yet remain underexplored in the important and challenging expressive talking head generation. In this work, we propose a DreamTalk framework to fulfill this gap, which employs meticulous design to unlock the potential of diffusion models in generating expressive talking heads. Specifically, DreamTalk consists of three crucial components: a denoising network, a style-aware lip expert, and a style predictor. The diffusion-based denoising network is able to consistently synthesize high-quality audio-driven face motions across diverse expressions. To enhance the expressiveness and accuracy of lip motions, we introduce a style-aware lip expert that can guide lip-sync while being mindful of the speaking styles. To eliminate the need for expression reference video or text, an extra diffusion-based style predictor is utilized to predict the target expression directly from the audio. By this means, DreamTalk can harness powerful diffusion models to generate expressive faces effectively and reduce the reliance on expensive style references. Experimental results demonstrate that DreamTalk is capable of generating photo-realistic talking faces with diverse speaking styles and achieving accurate lip motions, surpassing existing state-of-the-art counterparts.

1. Introduction

Audio-driven talking head generation, which concerns animating portraits with speech audio, has garnered significant interest due to its diverse applications in video games, film dubbing, and virtual avatars. Generating lifelike facial expressions is essential for enhancing the realism of talking heads [19]. These expressions during speech are termed as speaking styles [12, 46]. GANs [21] currently hold the state-of-the-art in expressive talking head generation [20, 30, 37, 46].However, their inherent issues with mode collapse and unstable training hamper their efficacy in consistently achieving high performance across a diverse range of speaking styles. Another issue is that prior methods often rely on reference videos [30, 37, 78] or texts [45, 90] to specify speaking styles. Their acquisition requires extra manual effort and hence is inflexible.

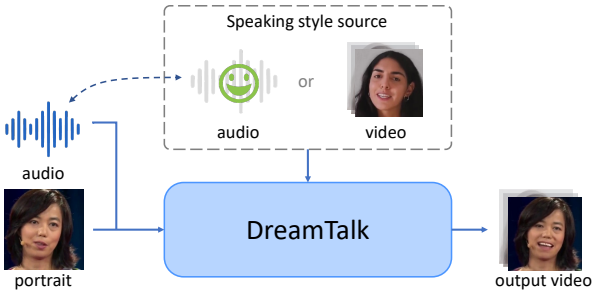

Figure 1. Leveraging the powerful diffusion models, DreamTalk is able to generate highly expressive talking heads across diverse speaking styles. Furthermore, DreamTalk is able to derive personalized speaking style directly from input audio, which obviates the need for additional style references.

As a new line of generative technique, diffusion models [26, 62] have recently been shown to produce high-quality results in numerous generative areas such as image generation [16, 53], video generation [27, 60, 86], and human motion synthesis [1, 72]. The success of diffusion models, stemming from their superior properties such as powerful distribution learning [16, 72], good convergence,

*Intern at Alibaba Group.

and stylistic diversity, make them exceptionally promising for exploring expressive talking head generation. However, current diffusion-based talking head approaches [58, 66, 96] primarily concentrate on generating talking heads with neutral expressions and still struggle to produce satisfactory performance, *e.g.*, suffering from frame jittering problem [58]. As a result, how to stimulate the full potential of diffusion models for expressive talking head generation is a promising yet untapped research direction.

In this paper, we propose DreamTalk, an expressive talking head generation framework that takes advantage of diffusion models to simultaneously deliver high performance across diverse speaking styles and reduce the reliance on expensive style references. Specifically, DreamTalk is composed of a denoising network, a style-aware lip expert, and a style predictor. The diffusion-based denoising network produces audio-driven facial motions with the speaking style specified by a reference video. The great distribution-learning characteristic of diffusion models endows the denoising network with the potential to produce high-quality results across diverse speaking styles. To harness this potential, we design a style-aware lip expert that drives the denoising network to produce accurate lip motions with vivid expressions. Contrasting with previous lip experts that overlook expression information and thus compromise style expressiveness, the proposed lip expert not only enhances lip accuracy but also ensures expressiveness. Finally, to further eliminate the need for additional style references, a diffusion-based style predictor is incorporated to predict personalized speaking styles directly from audio. The style predictor also incorporates the portrait as input during prediction, leveraging the correlation between speaker identity and speaking styles, thereby enhancing performance.

Ultimately, DreamTalk can consistently generate photo-realistic talking faces with precise lip-sync across a wide range of speaking styles while minimizing the need for additional style references. It also enables versatile manipulation of speaking styles and exhibits robust generalization across varied inputs, including songs, speech in multiple languages, noisy audio, and out-of-domain portraits. The effectiveness of DreamTalk is demonstrated through comprehensive qualitative and quantitative evaluations, showcasing its superiority over existing state-of-the-art methods.

2. Related Work

Audio-Driven Talking Head Generation. Audio-driven methods [14, 22, 67, 80, 81, 102] fall into two main categories: person-specific and person-agnostic. Person-specific approaches [18, 29, 41, 68, 82] are constrained to generating videos for speakers seen during training. Many of these [29, 34, 35, 64, 70, 74, 94, 97, 98] first craft 3D facial animations, later converting them into realistic videos. Recent advancements [23, 39, 57, 71, 71, 93] have employed neural radiance fields for modeling, yielding high-fidelity, realistic videos. Conversely, person-agnostic methods [10, 55, 77, 84] target generating videos for unseen speakers. Early methods prioritized lip synchronization [5, 6, 49, 65, 77, 103]. Later works shifted focus to natural facial expressions [95, 105] and head poses [7, 83, 100, 101, 104].

Expressive Talking Head Generation. Early methods [13, 20, 24, 29, 61, 69, 82, 90] model expressions in discrete emotion classes. To model more fine-grained expressions, most recent methods [30, 37, 46] leverage an expression reference video and transfer the expressions from that video to the generated one. However, these GAN-based models suffer from mode collapse, leading to videos with inferior lip-sync and style expressiveness. Our work addresses these issues by using diffusion models.

Specifying desired speaking styles effortlessly is also important for users. Most previous methods specify speaking styles using reference videos [30, 37, 46] or text [20, 45, 90], which needs human labor. A more user-friendly approach is to derive speaking styles from the input audio. Previous methods can only infer a limited number of discrete emotion classes from audio signals [29, 61, 90]. TH-PAD [95] generates expressions only aligned with the audio rhythm, not aligning with the emotional content of the audio. Besides, previous methods neglect the information in the input portrait. In this work, we aim to infer personalized and emotional expressions using input audio and portraits.

Diffusion Models. Diffusion models [26, 62] have demonstrated strong performance across multiple vision tasks [16, 53, 86, 99], including text-to-image generation [31, 54], image inpainting [33, 42], human motion generation [1, 72], 3D content generation [43], and video generation [50, 60, 79, 87, 89]. Previous efforts [3, 47, 58, 66] employing diffusion models in talking head generation only produce talking heads with neutral emotion and the results are unsatisfactory. Some methods devise diffusion-based renderers [17, 91] or face motion prior [96], yet still use GAN or regression models to model the audio-motion mapping. In this work, we endeavor to harness diffusion models for the generation of expressive talking heads, presenting a more intricate challenge with greater practical relevance.

3. Method

3.1. Problem Formulation

Given a portrait $\boldsymbol{I}$, a speech $\boldsymbol{A}$, and a style reference video $\boldsymbol{R}$, our method aims to generate a talking head video with lip motions synchronized with the speech and the speaking style reflected in the reference video. The audio $\boldsymbol{A} = [\boldsymbol{a}_i]_{i=1}^{L}$ is parameterized as a sequence of acoustic features. $\boldsymbol{R}$ is a sequence of video frames.

Besides, to eliminate the need for extra style references,

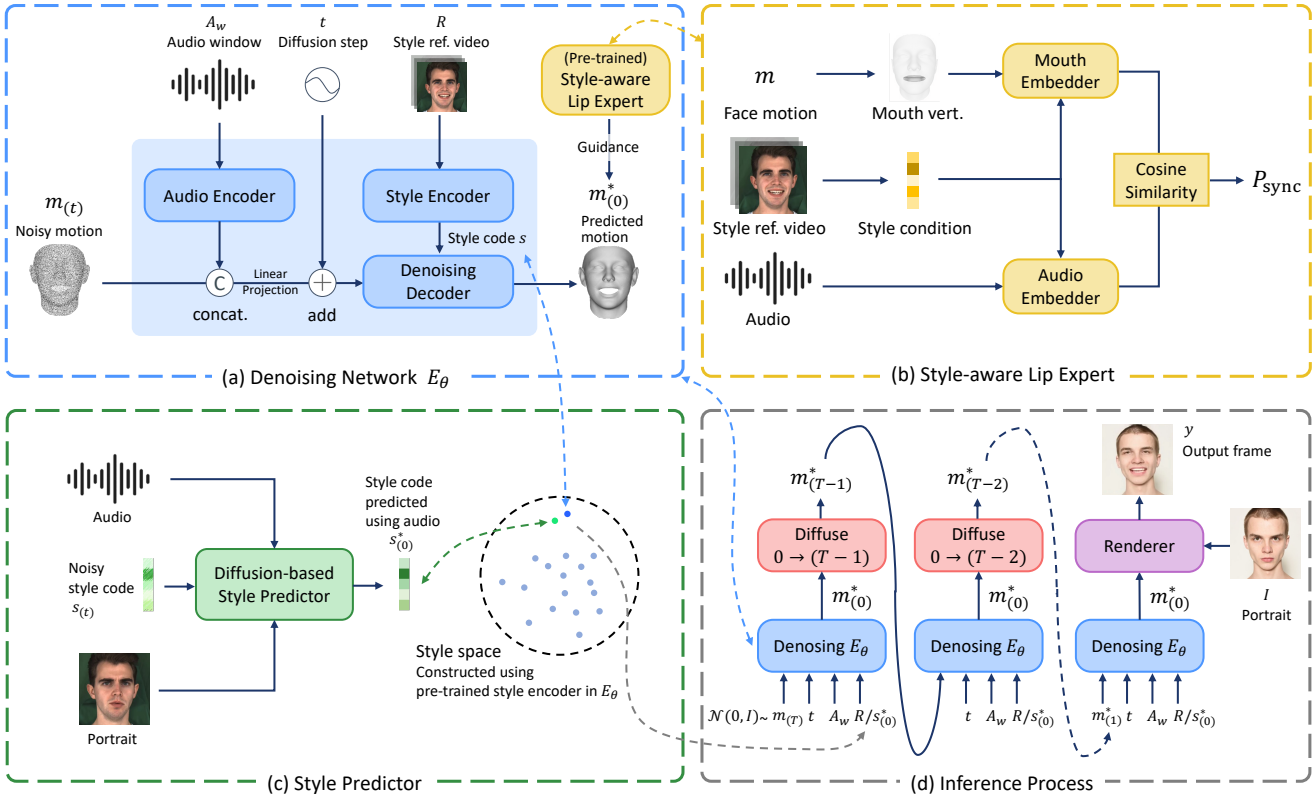

Figure 2. Illustration of DreamTalk. A style-aware lip expert (b), which evaluates the lip-audio synchronous probability under a given speaking style, is first trained to provide lip motion guidance for the denoising network (a). The denoising network is then trained to take the audio, the style reference video, and noisy face motion as input and predict the unnoised face motion. Then, A style predictor (c) is trained to predict the style code extracted from a video, taking audio and the speaker in that video as input. During inference process (d), the speaking style can be specified using style codes that are extracted from videos or derived from audio.

our method also aims to infer the speaking style using solely the speech and the portrait. The inferred speaking style can replace the role of style reference videos in controlling the expressions, which enables our method to generate expressive talking head videos with only speech and portrait input.

3.2. DreamTalk

DreamTalk comprises 3 key components: a denoising network, a style-aware lip expert, and a style predictor.

The denoising network computes face motion conditioned on the speech and style reference video. The face motion $\boldsymbol{M} = [\boldsymbol{m}_l]_{l=1}^{L}$ is parameterized as a sequence of expression parameters from 3D Morphable Models [4]. The face motion is rendered into video frames by a renderer [52]. The style-aware lip expert provides lip motion guidance under diverse expressions and thus drives the denoising network to achieve accurate lip-sync while ensuring style expressiveness. The style predictor can predict the speaking style aligned with that conveyed in speech.

Denoising Network. The denoising network synthesizes face motion sequence frame-by-frame in a sliding window manner. It predicts a motion frame $\boldsymbol{m}_l$ using an audio window $\boldsymbol{A}_w = [\boldsymbol{a}_i]_{i=l-w}^{l+w}$, where w denotes the window size.

The denoising network leverages forward and reverse diffusion processes. The diffusion process is modeled as a Markov noising process. Starting from a motion frame $\boldsymbol{m}_{(0)}$, it incrementally introduces Gaussian noise into the real data, gradually diffusing towards a distribution resembling $\mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$. Consequently, the distribution evolves as follows:

$$q(\boldsymbol{m}_{(t)}|\boldsymbol{m}_{(t-1)}) = \mathcal{N}(\sqrt{\alpha_n}\boldsymbol{m}_{(t-1)}, (1-\alpha_n)\boldsymbol{I}), \quad (1)$$

where $\boldsymbol{m}_{(t)}$ is the motion frame sampled at diffusion step t, $t \in \{1, \ldots, T\}$, and α_n is determined by the variance schedules. Conversely, the reverse diffusion process, or the denoising process, predicts the added noise in a noisy motion frame. Starting from a random motion frame $\boldsymbol{m}_{(T)} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$, the denoising process incrementally removes the noise and recovers the original motion $\boldsymbol{m}_{(0)}$.

Instead of predicting the noise as formulated by [26], we follow [51] and predict the signal itself. The denoising network E_θ predicts $\boldsymbol{m}_{(0)}$ based on the noisy motion, the diffusion step, the speech context, and the style reference:

$$\boldsymbol{m}^*_{(0)} = E_\theta(\boldsymbol{m}_{(t)}, t, \boldsymbol{A}_w, \boldsymbol{R}). \quad (2)$$

The asterisk($*$) indicates quantities that are generated.

Our denoising network has a transformer architecture [76]. The audio window $\boldsymbol{A}_w$ is first fed into a transformer-based audio encoder and the output is concatenated with the noisy motion $\boldsymbol{m}_{(t)}$ in the channel dimension. After linearly projected to the same dimension, the concatenated results and the timestep t are summed and served as the key and value of a transformer decoder. To extract the speaking style from the style reference, a style encoder first extracts the sequence of 3DMM expression parameters from $\boldsymbol{R}$ and then feeds them into a transformer encoder. The output tokens are aggregated using a self-attention pooling layer [56] to obtain the style code $\boldsymbol{s}$. The style code is repeated $2w+1$ times and added with positional encodings. The results serve as the query of the transformer decoder. The middle output token of the decoder is fed into a feed-forward network to predict the signal $\boldsymbol{m}_{(0)}$.

Style-aware Lip Expert. We observe that using solely the denoising loss in standard diffusion models results in inaccurate lip motions. We conjecture that the loss alone is insufficient for the denoising network to effectively focus on generating precise lip motions. A typical remedy is to involve a pre-trained lip expert [49] that provides lip motion guidance. However, we observe the lip expert reduces the intensity of expressions. This stems from the fact that the lip expert merely focuses on a generic speaking style, which leads to generating face motions in a uniform style.

To address this issue, we introduce a style-aware lip expert. The proposed lip expert is trained to evaluate lip-sync under diverse speaking styles. Therefore, it can provide lip motion guidance under diverse speaking styles and strike a better balance between style expressiveness and lip-sync. The lip expert $\mathcal{E}$ computes the probability that a clip of audio and lip motions are synchronous conditioned on style reference $\boldsymbol{R}$:

$$P_{\text{sync}} = \mathcal{E}([\boldsymbol{a}_i]_{i=l}^{l+n}, [\boldsymbol{m}_i]_{i=l}^{l+n}, \boldsymbol{R}), \tag{3}$$

where n denotes the clip length.

The style-aware lip expert encodes the lip motions and audio into respective embeddings conditioned on style reference and then computes the cosine similarity to represent the sync probability. To obtain lip motion information from face motion $\boldsymbol{m}$, we first convert $\boldsymbol{m}$ into the corresponding face mesh and select vertices in the mouth area as the representation of the lip motion [46]. The lip motion and audio encoders are mainly implemented by MLPs and 1D-convolutions, respectively. The style condition is fused into embeddings by first extracting style features from style reference using a style encoder, which mirrors the architecture of the one in the denoising network, and then concatenating the style features with intermediate feature maps from embedding encoders. The style encoder in the lip expert and the generator do not share parameters.

Style Predictor. Specifically, the style predictor S_ϕ predicts the style code $\boldsymbol{s}$ extracted by the style encoder in the trained denoising network. Observing the correlation between speaker identity and style codes (Sec. 4.4), the style predictor also integrates the portrait as input. The style predictor is instantiated as a diffusion model and is trained to predict the style code itself:

$$\boldsymbol{s}^*_{(0)} = S_\phi(\boldsymbol{s}_{(t)}, t, \boldsymbol{A}, \boldsymbol{I}), \tag{4}$$

where $\boldsymbol{s}_{(t)}$ is the style code sampled at diffusion step t.

The style predictor S_ϕ is a transformer encoder on a sequence consisting of, in order: audio embeddings, an embedding for the diffusion timestep, a speaker info embedding, the noised style code embedding, and a final embedding called learned query whose output is used to predict the unnoised style code. Audio embeddings are audio features extracted using self-supervised pre-trained speech models. To obtain the speaker info embedding, our method first extracts the 3DMM identity parameters, which include the face shape information but removes irrelevant information, such as expressions, from the portrait, and then embeds it into a token using an MLP.

3.3. Training and Inference

Training. The style-aware lip expert is first pre-trained by determining whether randomly sampled audio and lip motion clips are synchronous as in [49] and then frozen during training the denoising network.

The denoising network E_θ is trained by sampling random tuples $(\boldsymbol{m}_{(0)}, t, \boldsymbol{A}_w, \boldsymbol{R})$ from dataset, corrupting $\boldsymbol{m}_{(0)}$ into $\boldsymbol{m}_{(t)}$ by adding Gaussian noises, executing denoising steps to $\boldsymbol{m}_{(t)}$, and optimizing the loss:

$$\mathcal{L}_{\text{net}} = \lambda_{\text{denoise}}\mathcal{L}_{\text{denoise}} + \lambda_{\text{sync}}\mathcal{L}_{\text{sync}}. \tag{5}$$

Specifically, the ground-truth motion $\boldsymbol{m}_{(0)}$, and the speech audio window $\boldsymbol{A}_w$ are extracted from the training video of the same moment. t is drawn from the uniform distribution $\mathcal{U}\{1, T\}$. The style reference $\boldsymbol{R}$ is a video clip randomly drawn from the same video containing $\boldsymbol{m}_{(0)}$.

We first compute the denoising loss of the diffusion models [26] defined as:

$$\mathcal{L}_{\text{denoise}} = \|\boldsymbol{m}_{(0)} - E_\theta(\boldsymbol{m}_{(t)}, t, \boldsymbol{A}_w, \boldsymbol{R})\|_2^2. \tag{6}$$

Then, the denoising network maximizes the synchronous probability via a sync loss on generated clips:

$$\mathcal{L}_{\text{sync}} = -\log(P_{\text{sync}}). \tag{7}$$

Classifier-free guidance [25] is used to train our model. Specifically, E_θ is trained to learn both the style-conditional and unconditional distributions via randomly setting $\boldsymbol{R} = \varnothing$ by 10% chance during training. $\varnothing$ is implemented as a

Methods	MEAD / HDTF / Voxceleb2				
	SSIM↑	CPBD↑	F-LMD↓	M-LMD↓	$\text{Sync}_{\text{conf}}$ ↑
MakeItTalk [105]	0.73 / 0.57 / 0.52	0.11 / 0.24 / 0.24	3.97 / 5.12 / 6.29	5.32 / 4.55 / 5.15	2.10 / 3.16 / 2.17
Wav2Lip [49]	0.80 / 0.63 / 0.54	**0.18** / 0.30 / **0.30**	2.72 / 4.53 / 5.85	4.05 / 3.60 / 4.64	**5.26** / **5.83** / **5.70**
PC-AVS [104]	0.50 / 0.42 / 0.36	0.07 / 0.13 / 0.09	5.83 / 9.71 / 12.9	4.97 / 4.17 / 7.42	2.18 / 4.85 / 4.73
AVCT [84]	0.83 / 0.76 / 0.64	0.14 / 0.22 / 0.23	2.92 / 2.86 / 3.62	5.52 / 3.57 / 3.71	2.53 / 4.27 / 3.89
GC-AVT [37]	0.34 / 0.36 / -	0.14 / 0.28 / -	8.04 / 10.2 / -	7.10 / 6.23 / -	2.42 / 4.72 / -
EAMM [30]	0.40 / 0.40 / 0.43	0.08 / 0.14 / 0.20	6.70 / 7.03 / 6.36	6.48 / 6.86 / 4.89	1.41 / 2.54 / 2.24
StyleTalk [46]	0.84 / 0.81 / 0.66	0.16 / 0.30 / 0.29	2.12 / 1.96 / 2.92	3.25 / 2.41 / 2.96	3.47 / 4.82 / 4.51
SadTalker [100]	0.69 / 0.77 / 0.44	0.16 / 0.24 / 0.19	4.12 / 5.99 / 9.12	4.37 / 4.07 / 6.11	2.76 / 4.35 / 4.38
PD-FGC [78]	0.49 / 0.41 / 0.35	0.05 / 0.13 / 0.12	5.50 / 9.50 / 12.5	4.10 / 4.23 / 8.19	2.27 / 4.68 / 4.64
EAT [20]	0.53 / 0.59 / 0.47	0.15 / 0.26 / 0.20	5.54 / 3.86 / 5.53	4.79 / 4.03 / 5.88	2.16 / 4.54 / 4.35
DreamTalk	**0.86** / **0.85** / **0.69**	0.16 / **0.31** / **0.30**	**1.93** / **1.80** / **2.69**	**2.91** / **2.15** / **2.72**	3.78 / 5.17 / 4.90
Ground Truth	1 / 1 / 1	0.22 / 0.31 / 0.33	0 / 0 / 0	0 / 0 / 0	4.13 / 5.44 / 5.23

Table 1. Quantitative comparisons on MEAD, HDTF, and Voxceleb2. Since we only receive GC-AVT samples on MEAD and HDTF, GC-AVT is not evaluated on Voxceleb2.

sequence of face motions $[\boldsymbol{m}_i]$ with all zero values. For inference, the predicted signal is computed by

$$\begin{aligned}\boldsymbol{m}^*_{(0)} = &\omega E_\theta(\boldsymbol{m}_{(t)}, t, \boldsymbol{A}_w, \boldsymbol{R}) \\ &+ (1-\omega) E_\theta(\boldsymbol{m}_{(t)}, t, \boldsymbol{A}_w, \varnothing),\end{aligned} \tag{8}$$

instead of Equation 2. This approach enables controlling the effectiveness of the style reference $\boldsymbol{R}$ through adjustment of the scale factor ω.

When training the style predictor, we draw a random video, then extract audio $\boldsymbol{A}$ and style code $\boldsymbol{s}_{(0)}$ (using the trained style encoder) from it. Since 3DMM identity parameters may leak expression information, the portrait $\boldsymbol{I}$ is sampled from another video with the same speaker identity. The style predictor E_ϕ is trained by optimizing the loss:

$$\mathcal{L}_{\text{pred}} = \|\boldsymbol{s}_{(0)} - S_\phi(\boldsymbol{s}_{(t)}, t, \boldsymbol{A}, \boldsymbol{I})\|_2^2, \tag{9}$$

We utilize PIRenderer [52] as the renderer and meticulously fine-tune it to empower the renderer with emotional expression generation capabilities.

Inference. Our method enables the specification of speaking styles using either reference videos or solely through input audio and portrait. In the case of reference videos, style codes are derived using the style encoder in the denoising network. When relying solely on input audio and portrait, these inputs are processed by the style predictor, which employs a denoising procedure to obtain the style code.

With the style code, the denoising network utilizes the sampling algorithm of DDPM [26] to produce face motions. It first samples a random motion $\boldsymbol{m}^*_{(T)} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$ then computes denoised sequences $\{\boldsymbol{m}^*_{(t)}\}, t = T-1, \ldots, 0$ by incrementally removing the noise from $\boldsymbol{m}^*_{(t)}$. Finally, the motion $\boldsymbol{m}^*_{(0)}$ is the generated face motion. The sampling process can be accelerated by leveraging DDIM [63]. The output face motions are then rendered into videos by the renderer PIRenderer.

4. Experiments

4.1. Experimental Setup

Datasets. We train and evaluate the denoising network on MEAD [82], HDTF [101], and Voxceleb2 [11]. Since Voxceleb2 official videos are of low resolution, we redownload the original YouTube videos and re-crop the videos. The style-aware lip expert is trained on MEAD and HDTF. We train the style predictor on MEAD and evaluate it on MEAD and RAVEDESS [40].

Baselines. We compare our method with previous methods including: MakeitTalk [105], Wav2Lip [49], PC-AVS [104], AVCT [84], GC-AVT [37], EAMM [30], StyleTalk [46], DiffTalk [58], SadTalker [100], PD-FGC [78], and EAT [20]. For DiffTalk, since the released model is incomplete and unable to generate reasonable results until submission, we perform qualitative comparisons using videos from its released demo. For other methods, we generate the samples using released models or with the help of the authors.

Metrics. We utilize widely used metrics: SSIM [88], the Cumulative Probability of Blur Detection (CPBD) [48], the SyncNet confidence score ($\text{Sync}_{\text{conf}}$) [9], the Landmark Distance around mouth area (M-LMD) [6], the Landmark Distance on the full face (F-LMD).

4.2. Main Results

Quanitative Comparisons. As shown in Tab. 1, our method outperforms previous methods across most metrics. Wav2Lip's training with SyncNet as a discriminator explains its high SyncNet confidence score, even surpassing the ground truth. Notably, our method's SyncNet confidence score closely aligns with the ground truth, and it achieves the best M-LMD scores, which indicates its capability for precise lip synchronization. Furthermore, our superior performance in the F-LMD metric demonstrates our

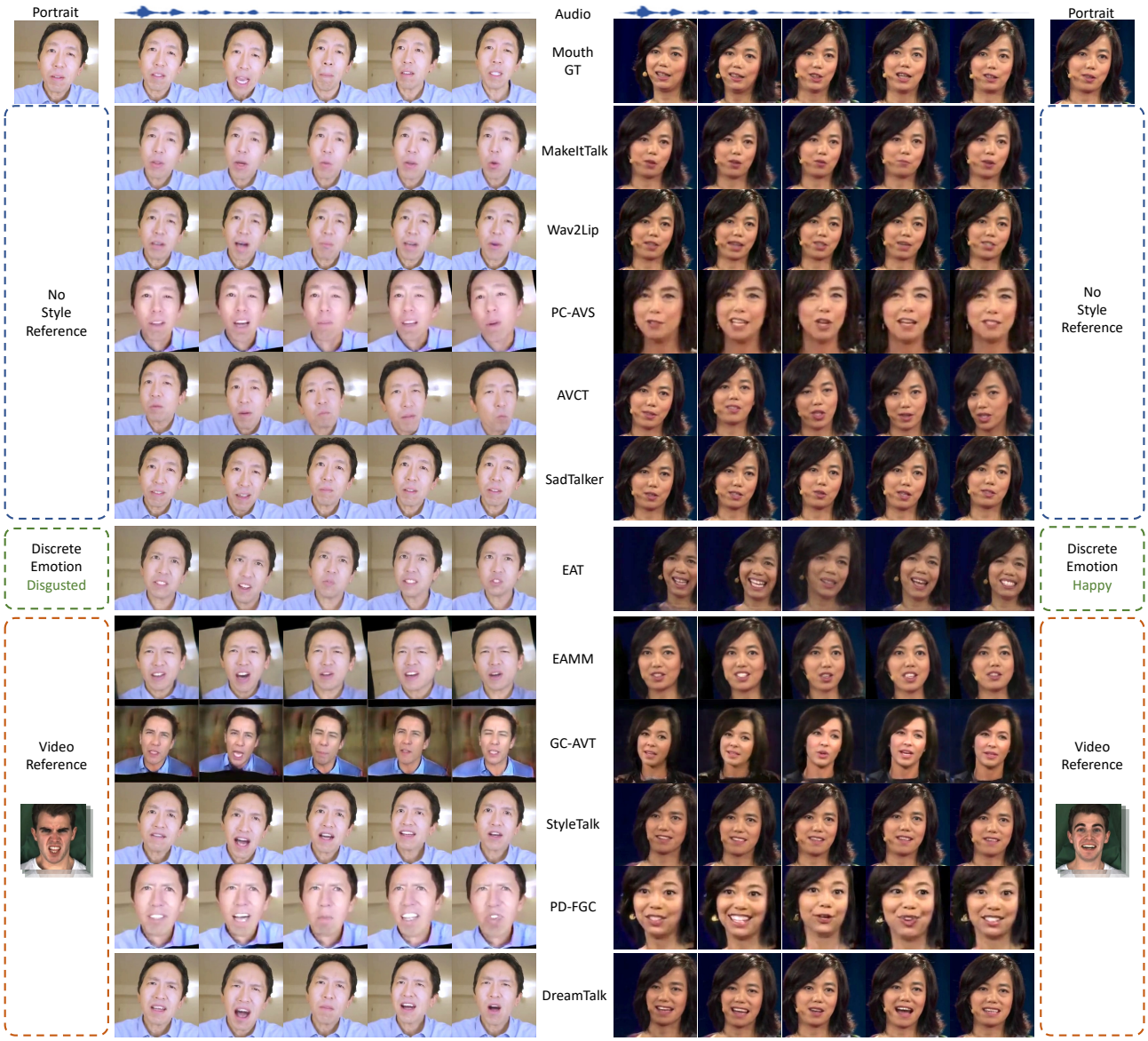

Figure 3. Qualitative comparisons with previous methods.

method's proficiency in generating facial expressions consistent with the reference speaking style.

Qualitative comparisons. Fig. 3 shows the qualitative comparisons. The portraits, style references, and audio are all unseen during training. It can be seen that MakeItTalk and AVCT struggle with accurate lip synchronization. While Wav2Lip and PC-AVS synchronize lips accurately, their outputs appear blurry. SadTalker, on the other hand, generally aligns lip movements with audio but occasionally displays unnatural jitters.

EAT's capability is limited to generating discrete emotions, lacking the finesse for nuanced expressions. For example, in the left case, the style reference shows the speaker narrowing his eyes, but EAT merely produces a generic disgusted look with wide-open glaring eyes. Additionally, as shown in the right case, EAT struggles to maintain a consistent face shape during speaker head movements.

EAMM, GC-AVT, StyleTalk, and PD-FGC demonstrate the ability to produce fine-grained expressions. However, EAMM falls short in lip synchronization, GC-AVT and PD-FGC struggle with preserving speaker identity, and all three have issues rendering a plausible background. We observed that StyleTalk, while capable of generating nuanced expressions, occasionally does so with diminished intensity and fails to generate accurate lip motion for some words. A notable example is shown in the third column of the left case: when the speaker utters "um"; the expected closed-mouth motion is replaced by an open mouth in StyleTalk's output.

Figure 4. Face distortion observed in StyleTalk's output.

Figure 5. Comparisons with DiffTalk.

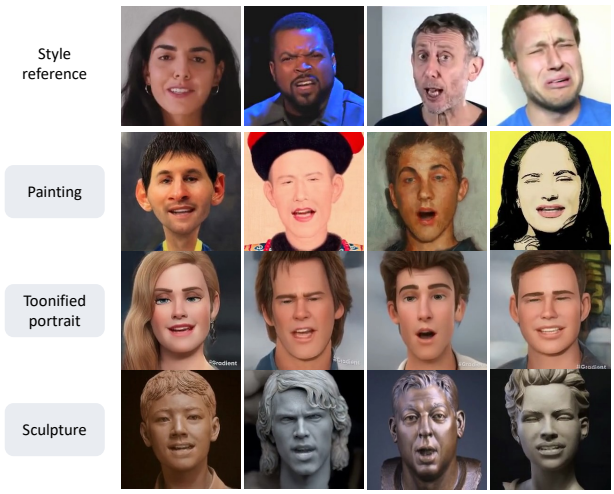

Figure 6. The results generated using out-of-domain portraits.

Besides, as shown in Fig. 4, StyleTalk occasionally generates videos where the face distorts suddenly.

Fig. 5 present comparisons with DiffTalk. DiffTalk struggles with lip synchronization and introduces jitteriness and artifacts in the mouth region.

In contrast, DreamTalk excels in producing realistic talking faces that not only mirror the reference speaking style but also achieve precise lip synchronization and superior video quality.

Generalization Capabilities. *Supplementary Video*

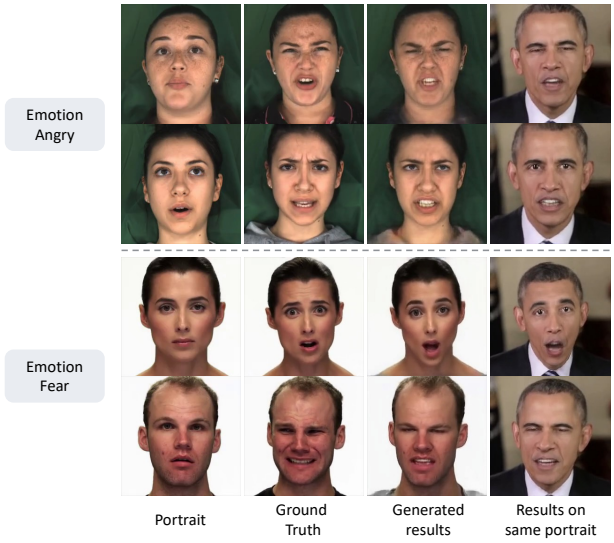

Figure 7. The results of speaking style prediction. The fourth column displays samples generated with predicted styles applied to the same portrait for clearer comparisons.

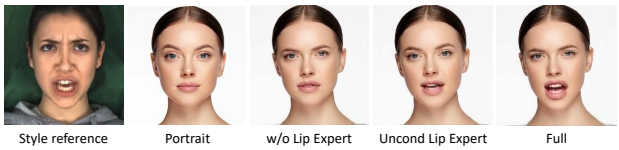

Figure 8. Ablation study results

demonstrates our method's capability to produce realistic talking head videos for out-of-domain portraits (paintings, toonified portraits, and sculptures, shown in Fig. 6), speech in various languages, noisy audio input, and songs. Appendix B.1 provides more analysis.

Results of Speaking Style Prediction. Fig. 7 presents the results of speaking style predictions. The speakers with angry emotion are sampled from MEAD while those with fear emotion are sampled from RAVEDESS. The style predictor, utilizing emotional audio and neutral portraits from test videos, adeptly deduces personalized speaking styles that align with those observed in the original videos. It demonstrates the capacity to discern subtle expressions within the same emotion. For instance, for samples with angry emotion, the first-row speaker exhibits narrowed eyes, in contrast to the second-row speaker's intense, glaring stare. For samples with angry emotion, the first-row speaker's eyes and mouth are open, whereas the second-row speaker combines narrowed eyes with a contorted facial expression.

The ablation study and user study for the style predictor are presented in Appendix A.

4.3. Ablation Study

To analyze the contributions of our designs, we conduct an ablation study with two variants: (1) remove the style-

Method	SSIM↑	F-LMD↓	M-LMD↓	$\text{Sync}_{\text{conf}}$↑
w/o Lip Expert	0.85	**1.90**	3.07	2.63
Uncond Lip Expert	0.83	2.19	3.42	**4.51**
Full	**0.86**	1.93	**2.91**	3.78

Table 2. The results of DreamTalk's ablation study on MEAD. We omit CPBD scores since there are no significant differences between variants on CPBD.

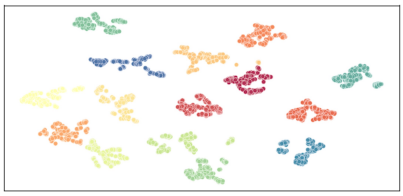

Figure 9. t-SNE visualization of style codes from 15 speakers. Each color stands for style codes from an identical speaker.

aware lip expert (**w/o lip expert**); (2) trained with unconditional lip expert (**uncond lip expert**). Our full model is denoted as **Full**.

Fig. 8 and Tab. 2 present our ablation study results. The variant **w/o lip expert** exhibits a decline in lip-sync accuracy on the emotional dataset MEAD, despite its competitive F-LMD score indicating expressive facial generation. Conversely, **uncond lip expert** secures a superior SyncNet confidence score at the expense of speaking style expressiveness. The **Full** model achieves a harmonious balance, ensuring both precise lip synchronization and vivid expressions, thanks to the style-aware lip expert directing the diffusion model's expressive potential.

4.4. Style Code Visualization

Using t-distributed stochastic neighbor embedding (t-SNE) [75], we map style codes from the MEAD dataset's 15 speakers into a 2D space. These speakers exhibit 22 distinct speaking styles, comprising seven emotions at three intensity levels, alongside a neutral style. For each style, style codes are extracted from 10 randomly chosen videos.

Fig. 9 reveals that style codes from identical speakers tend to cluster, despite some codes sharing identical emotion categories. This suggests that the variance in speaking styles attributable to individual speaker idiosyncrasies outweighs that induced by different emotions, thereby underpinning the rationale for using portrait information to infer speaking styles. We also observe that each speaker's style code distribution exhibits both common patterns and individualized characteristics. We present these intriguing observations in Appendix B.2.

4.5. Style Manipulation

Adjusting the scale of Classifer-free Guidance. As elaborated in Sec. 3.3, the scale factor ω in the classifier-free guidance scheme modulates the effect of the input style. Adjusting ω either amplifies or attenuates the designated

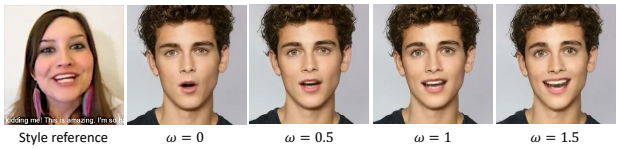

Figure 10. Effect of the input style controlled by adjusting the scale ω of the classifier-free guidance.

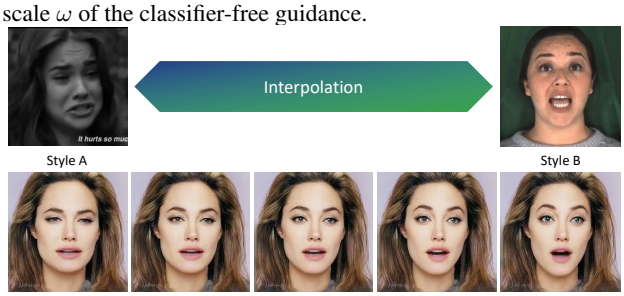

Figure 11. The results of speaking style interpolation.

Methods	Lip Sync↑	Realness↑	Style Consistency↑
MakeItTalk [105]	1.94	2.03	1.65
Wav2Lip [49]	2.29	1.45	1.14
PC-AVS [104]	2.26	1.81	1.86
AVCT [84]	2.31	2.21	1.72
GC-AVT [37]	2.35	1.20	1.88
EAMM [30]	1.81	1.40	1.78
StyleTalk [46]	2.35	2.29	2.08
SadTalker [100]	2.37	2.38	1.73
PD-FGC [78]	1.95	1.61	2.31
EAT [20]	2.24	1.65	2.29
DreamTalk	2.55	2.60	2.46
Ground Truth	3.03	2.89	1.83

Table 3. User study results.

style, as shown in Fig. 10. When $\omega = 0$, DreamTalk produces a talking head with a neutral expression. We observed that when the scale factor ω exceeds 2, there is a noticeable decline in lip-sync accuracy.

Style Code Interpolation. Leveraging the style space, we can modify speaking styles via style code manipulation. Fig. 11 illustrates that linear interpolation between style codes results in a seamless transition of generated speaking styles. This interpolation process allows for style intensity modulation and the generation of novel speaking styles.

4.6. User Study

We conduct a user study of 20 participants. We generate the test samples covering multiple speaking styles and speakers. For each method, the participant is required to score 10 videos sampled from the test samples and is asked to give a rating (from 1 to 5, 5 is the best) on three aspects: (1) the Lip sync quality, (2) the realness of results, and (3) the style consistency between the generated videos and the style reference. As shown in Tab. 3, our method outperforms existing approaches across all aspects, particularly in style consistency, highlighting its superior capabilities.

5. Conclusion

In this work, we propose DreamTalk, a novel approach leveraging diffusion models for generating expressive talking heads. Our method aims to excel in diverse speaking styles while minimizing dependence on extra style references. We develop a denoising network for creating expressive, audio-driven facial motions and introduce a style-aware lip expert to optimize lip-sync without compromising style expressiveness. Additionally, we devise a style predictor that infers speaking styles directly from audio, eliminating the need for video references. The efficacy of DreamTalk is validated through extensive experiments.

Acknowledgements. This work is supported by Alibaba Group through Alibaba Research Intern Program. We would like to thank Xinya Ji, Borong Liang, Yan Pan, and Suzhen Wang for their generous help with the comparisons.

References

[1] Tenglong Ao, Zeyi Zhang, and Libin Liu. Gesturediffuclip: Gesture diffusion model with clip latents. *arXiv preprint arXiv:2303.14613*, 2023. 1, 2

[2] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460, 2020. 16

[3] Dan Bigioi, Shubhajit Basak, Hugh Jordan, Rachel McDonnell, and Peter Corcoran. Speech driven video editing via an audio-conditioned diffusion model. *arXiv preprint arXiv:2301.04474*, 2023. 2

[4] 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. 3

[5] Lele Chen, Zhiheng Li, Ross K Maddox, Zhiyao Duan, and Chenliang Xu. Lip movements generation at a glance. In *ECCV*, pages 520–535, 2018. 2

[6] Lele Chen, Ross K Maddox, Zhiyao Duan, and Chenliang Xu. Hierarchical cross-modal talking face generation with dynamic pixel-wise loss. In *CVPR*, pages 7832–7841, 2019. 2, 5

[7] Lele Chen, Guofeng Cui, Celong Liu, Zhong Li, Ziyi Kou, Yi Xu, and Chenliang Xu. Talking-head generation with rhythmic head motion. In *ECCV*, pages 35–51. Springer, 2020. 2

[8] Weidong Chen, Xiaofeng Xing, Xiangmin Xu, Jichen Yang, and Jianxin Pang. Key-sparse transformer for multimodal speech emotion recognition. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6897–6901. IEEE, 2022. 17

[9] Joon Son Chung and Andrew Zisserman. Out of time: automated lip sync in the wild. In *Asian conference on computer vision*, pages 251–263. Springer, 2016. 5

[10] Joon Son Chung, Amir Jamaludin, and Andrew Zisserman. You said that? *arXiv preprint arXiv:1705.02966*, 2017. 2

[11] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. Voxceleb2: Deep speaker recognition. *arXiv preprint arXiv:1806.05622*, 2018. 5

[12] Daniel Cudeiro, Timo Bolkart, Cassidy Laidlaw, Anurag Ranjan, and Michael J Black. Capture, learning, and synthesis of 3d speaking styles. In *CVPR*, pages 10101–10111, 2019. 1

[13] Radek Daněček, Kiran Chhatre, Shashank Tripathi, Yandong Wen, Michael J Black, and Timo Bolkart. Emotional speech-driven animation with content-emotion disentanglement. *arXiv preprint arXiv:2306.08990*, 2023. 2

[14] Dipanjan Das, Sandika Biswas, Sanjana Sinha, and Brojeshwar Bhowmick. Speech-driven facial animation using cascaded gans for learning of motion and texture. In *ECCV*, pages 408–424. Springer, 2020. 2

[15] 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 *CVPRW*, pages 0–0, 2019. 15

[16] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *NeurIPS*, 2021. 1, 2

[17] Chenpng Du, Qi Chen, Tianyu He, Xu Tan, Xie Chen, Kai Yu, Sheng Zhao, and Jiang Bian. Dae-talker: High fidelity speech-driven talking face generation with diffusion autoencoder. *arXiv preprint arXiv:2303.17550*, 2023. 2

[18] Ohad Fried, Ayush Tewari, Michael Zollhöfer, Adam Finkelstein, Eli Shechtman, Dan B Goldman, Kyle Genova, Zeyu Jin, Christian Theobalt, and Maneesh Agrawala. Text-based editing of talking-head video. *ACM Transactions on Graphics (TOG)*, 38(4):1–14, 2019. 2

[19] Chris Frith. Role of facial expressions in social interactions. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1535):3453–3458, 2009. 1

[20] 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, 2, 5, 8, 16

[21] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014. 1

[22] Jiazhi Guan, Zhanwang Zhang, Hang Zhou, Tianshu Hu, Kaisiyuan Wang, Dongliang He, Haocheng Feng, Jingtuo Liu, Errui Ding, Ziwei Liu, et al. Stylesync: High-fidelity generalized and personalized lip sync in style-based generator. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1505–1515, 2023. 2

[23] Yudong Guo, Keyu Chen, Sen Liang, Yongjin Liu, Hujun Bao, and Juyong Zhang. Ad-nerf: Audio driven neural radiance fields for talking head synthesis. *arXiv preprint arXiv:2103.11078*, 2021. 2

[24] Siddharth Gururani, Arun Mallya, Ting-Chun Wang, Rafael Valle, and Ming-Yu Liu. Space: Speech-driven portrait animation with controllable expression. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20914–20923, 2023. 2

[25] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021. 4

[26] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, pages 6840–6851. Curran Associates, Inc., 2020. 1, 2, 3, 4, 5

[27] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022. 1

[28] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 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. 16

[29] Xinya Ji, Hang Zhou, Kaisiyuan Wang, Wayne Wu, Chen Change Loy, Xun Cao, and Feng Xu. Audio-driven emotional video portraits. In *CVPR*, pages 14080–14089, 2021. 2

[30] 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. *arXiv preprint arXiv:2205.15278*, 2022. 1, 2, 5, 8, 15, 16

[31] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. *arXiv*, 2022. 2

[32] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 16

[33] Wing-Fung Ku, Wan-Chi Siu, Xi Cheng, and H Anthony Chan. Intelligent painter: Picture composition with resampling diffusion model. *arXiv*, 2022. 2

[34] Avisek Lahiri, Vivek Kwatra, Christian Frueh, John Lewis, and Chris Bregler. Lipsync3d: Data-efficient learning of personalized 3d talking faces from video using pose and lighting normalization. In *CVPR*, pages 2755–2764, 2021. 2

[35] Lincheng Li, Suzhen Wang, Zhimeng Zhang, Yu Ding, Yixing Zheng, Xin Yu, and Changjie Fan. Write-a-speaker: Text-based emotional and rhythmic talking-head generation. In *AAAI*, pages 1911–1920, 2021. 2

[36] Yongyuan Li, Xiuyuan Qin, Chao Liang, and Mingqiang Wei. Hdtr-net: A real-time high-definition teeth restoration network for arbitrary talking face generation methods. *arXiv preprint arXiv:2309.07495*, 2023. 16

[37] 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 *CVPR*, pages 3387–3396, 2022. 1, 2, 5, 8

[38] Pengfei Liu, Wenjin Deng, Hengda Li, Jintai Wang, Yinglin Zheng, Yiwei Ding, Xiaohu Guo, and Ming Zeng. Musicface: Music-driven expressive singing face synthesis. *arXiv preprint arXiv:2303.14044*, 2023. 14

[39] Xian Liu, Yinghao Xu, Qianyi Wu, Hang Zhou, Wayne Wu, and Bolei Zhou. Semantic-aware implicit neural audio-driven video portrait generation. *arXiv preprint arXiv:2201.07786*, 2022. 2

[40] Steven R Livingstone and Frank A Russo. The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english. *PloS one*, 13(5): e0196391, 2018. 5

[41] Yuanxun Lu, Jinxiang Chai, and Xun Cao. Live speech portraits: real-time photorealistic talking-head animation. *ACM Transactions on Graphics (TOG)*, 40(6):1–17, 2021. 2

[42] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *CVPR*, 2022. 2

[43] Shitong Luo and Wei Hu. Diffusion probabilistic models for 3d point cloud generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2837–2845, 2021. 2

[44] Tian Lv, Yu-Hui Wen, Zhiyao Sun, Zipeng Ye, and Yong-Jin Liu. Generating smooth and facial-details-enhanced talking head video: A perspective of pre and post processes. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 7079–7083, 2022. 16

[45] 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. 1, 2

[46] 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*, 2023. 1, 2, 4, 5, 8, 14, 15, 16

[47] Soumik Mukhopadhyay, Saksham Suri, Ravi Teja Gadde, and Abhinav Shrivastava. Diff2lip: Audio conditioned diffusion models for lip-synchronization. *arXiv preprint arXiv:2308.09716*, 2023. 2

[48] Niranjan D Narvekar and Lina J Karam. A no-reference perceptual image sharpness metric based on a cumulative probability of blur detection. In *2009 International Workshop on Quality of Multimedia Experience*, pages 87–91. IEEE, 2009. 5

[49] 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, 4, 5, 8

[50] Zhiwu Qing, Shiwei Zhang, Jiayu Wang, Xiang Wang, Yujie Wei, Yingya Zhang, Changxin Gao, and Nong Sang. Hierarchical spatio-temporal decoupling for text-to-video generation. *arXiv preprint arXiv:2312.04483*, 2023. 2

[51] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022. 3

[52] Yurui Ren, Ge Li, Yuanqi Chen, Thomas H Li, and Shan Liu. Pirenderer: Controllable portrait image generation via semantic neural rendering. In *ICCV*, pages 13759–13768, 2021. 3, 5, 16

[53] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 1, 2

[54] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. *arXiv*, 2022. 2

[55] Najmeh Sadoughi and Carlos Busso. Speech-driven expressive talking lips with conditional sequential generative adversarial networks. *IEEE Transactions on Affective Computing*, 12(4):1031–1044, 2019. 2

[56] Pooyan Safari, Miquel India, and Javier Hernando. Self-attention encoding and pooling for speaker recognition. *arXiv preprint arXiv:2008.01077*, 2020. 4, 15

[57] Shuai Shen, Wanhua Li, Zheng Zhu, Yueqi Duan, Jie Zhou, and Jiwen Lu. Learning dynamic facial radiance fields for few-shot talking head synthesis. In *European Conference on Computer Vision*, pages 666–682. Springer, 2022. 2

[58] Shuai Shen, Wenliang Zhao, Zibin Meng, Wanhua Li, Zheng Zhu, Jie Zhou, and Jiwen Lu. Difftalk: Crafting diffusion models for generalized talking head synthesis. *arXiv preprint arXiv:2301.03786*, 2023. 2, 5

[59] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. First order motion model for image animation. *Advances in Neural Information Processing Systems*, 32:7137–7147, 2019. 15

[60] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022. 1, 2

[61] Sanjana Sinha, Sandika Biswas, Ravindra Yadav, and Brojeshwar Bhowmick. Emotion-controllable generalized talking face generation. *arXiv preprint arXiv:2205.01155*, 2022. 2

[62] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *ICML*, 2015. 1, 2

[63] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 5

[64] Linsen Song, Wayne Wu, Chen Qian, Ran He, and Chen Change Loy. Everybody's talkin': Let me talk as you want. *arXiv preprint arXiv:2001.05201*, 2020. 2

[65] Yang Song, Jingwen Zhu, Dawei Li, Xiaolong Wang, and Hairong Qi. Talking face generation by conditional recurrent adversarial network. *arXiv preprint arXiv:1804.04786*, 2018. 2

[66] Michał Stypułkowski, Konstantinos Vougioukas, Sen He, Maciej Zięba, Stavros Petridis, and Maja Pantic. Diffused heads: Diffusion models beat gans on talking-face generation. *arXiv preprint arXiv:2301.03396*, 2023. 2

[67] Yasheng Sun, Hang Zhou, Kaisiyuan Wang, Qianyi Wu, Zhibin Hong, Jingtuo Liu, Errui Ding, Jingdong Wang, Ziwei Liu, and Koike Hideki. Masked lip-sync prediction by audio-visual contextual exploitation in transformers. In *SIGGRAPH Asia 2022 Conference Papers*, pages 1–9, 2022. 2

[68] Supasorn Suwajanakorn, Steven M Seitz, and Ira Kemelmacher-Shlizerman. Synthesizing obama: learning lip sync from audio. *ACM Transactions on Graphics (ToG)*, 36(4):1–13, 2017. 2

[69] 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. 2

[70] Anni Tang, Tianyu He, Xu Tan, Jun Ling, Runnan Li, Sheng Zhao, Li Song, and Jiang Bian. Memories are one-to-many mapping alleviators in talking face generation. *arXiv preprint arXiv:2212.05005*, 2022. 2

[71] Jiaxiang Tang, Kaisiyuan Wang, Hang Zhou, Xiaokang Chen, Dongliang He, Tianshu Hu, Jingtuo Liu, Gang Zeng, and Jingdong Wang. Real-time neural radiance talking portrait synthesis via audio-spatial decomposition. *arXiv preprint arXiv:2211.12368*, 2022. 2

[72] Guy Tevet, Sigal Raab, Brian Gordon, Yoni Shafir, Daniel Cohen-or, and Amit Haim Bermano. Human motion diffusion model. In *The Eleventh International Conference on Learning Representations*, 2023. 1, 2

[73] Joachim Thiemann, Nobutaka Ito, and Emmanuel Vincent. The diverse environments multi-channel acoustic noise database (demand): A database of multichannel environmental noise recordings. In *Proceedings of Meetings on Acoustics*. AIP Publishing, 2013. 14

[74] Justus Thies, Mohamed Elgharib, Ayush Tewari, Christian Theobalt, and Matthias Nießner. Neural voice puppetry: Audio-driven facial reenactment. In *ECCV*, pages 716–731. Springer, 2020. 2

[75] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9 (11), 2008. 8

[76] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2017. 4

[77] Konstantinos Vougioukas, Stavros Petridis, and Maja Pantic. Realistic speech-driven facial animation with gans. *International Journal of Computer Vision*, pages 1–16, 2019. 2

[78] 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, 5, 8

[79] Jiuniu Wang, Hangjie Yuan, Dayou Chen, Yingya Zhang, Xiang Wang, and Shiwei Zhang. Modelscope text-to-video technical report. *arXiv preprint arXiv:2308.06571*, 2023. 2

[80] Jiayu Wang, Kang Zhao, Yifeng Ma, Shiwei Zhang, Yingya Zhang, Yujun Shen, Deli Zhao, and Jingren Zhou. Facecomposer: A unified model for versatile facial content creation. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 2

[81] Jiayu Wang, Kang Zhao, Shiwei Zhang, Yingya Zhang, Yujun Shen, Deli Zhao, and Jingren Zhou. Lipformer: High-fidelity and generalizable talking face generation with a pre-learned facial codebook. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13844–13853, 2023. 2

[82] 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 *ECCV*, pages 700–717. Springer, 2020. 2, 5

[83] Suzhen Wang, Lincheng Li, Yu Ding, Changjie Fan, and Xin Yu. Audio2head: Audio-driven one-shot talking-head generation with natural head motion. *IJCAI*, 2021. 2, 16

[84] Suzhen Wang, Lincheng Li, Yu Ding, and Xin Yu. One-shot talking face generation from single-speaker audio-visual correlation learning. In *AAAI*, 2022. 2, 5, 8

[85] Suzhen Wang, Yifeng Ma, and Yu Ding. Exploring complementary features in multi-modal speech emotion recognition. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023. 17

[86] Xiang Wang, Hangjie Yuan, Shiwei Zhang, Dayou Chen, Jiuniu Wang, Yingya Zhang, Yujun Shen, Deli Zhao, and Jingren Zhou. Videocomposer: Compositional video synthesis with motion controllability. In *NeurIPS*, 2023. 1, 2

[87] Xiang Wang, Shiwei Zhang, Han Zhang, Yu Liu, Yingya Zhang, Changxin Gao, and Nong Sang. Videolcm: Video latent consistency model. *arXiv preprint arXiv:2312.09109*, 2023. 2

[88] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 5

[89] Yujie Wei, Shiwei Zhang, Zhiwu Qing, Hangjie Yuan, Zhiheng Liu, Yu Liu, Yingya Zhang, Jingren Zhou, and Hongming Shan. Dreamvideo: Composing your dream videos with customized subject and motion. *arXiv preprint arXiv:2312.04433*, 2023. 2

[90] Chao Xu, Junwei Zhu, Jiangning Zhang, Yue Han, Wenqing Chu, Ying Tai, Chengjie Wang, Zhifeng Xie, and Yong Liu. High-fidelity generalized emotional talking face generation with multi-modal emotion space learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6609–6619, 2023. 1, 2, 16

[91] Chao Xu, Shaoting Zhu, Junwei Zhu, Tianxin Huang, Jiangning Zhang, Ying Tai, and Yong Liu. Multimodal-driven talking face generation, face swapping, diffusion model. *arXiv preprint arXiv:2305.02594*, 2023. 2

[92] Zipeng Ye, Zhiyao Sun, Yu-Hui Wen, Yanan Sun, Tian Lv, Ran Yi, and Yong-Jin Liu. Dynamic neural textures: Generating talking-face videos with continuously controllable expressions. *arXiv preprint arXiv:2204.06180*, 2022. 16

[93] Zhenhui Ye, Ziyue Jiang, Yi Ren, Jinglin Liu, JinZheng He, and Zhou Zhao. Geneface: Generalized and high-fidelity audio-driven 3d talking face synthesis. *arXiv preprint arXiv:2301.13430*, 2023. 2

[94] Ran Yi, Zipeng Ye, Juyong Zhang, Hujun Bao, and Yong-Jin Liu. Audio-driven talking face video generation with learning-based personalized head pose. *arXiv preprint arXiv:2002.10137*, 2020. 2

[95] Zhentao Yu, Zixin Yin, Deyu Zhou, Duomin Wang, Finn Wong, and Baoyuan Wang. Talking head generation with probabilistic audio-to-visual diffusion priors. *arXiv preprint arXiv:2212.04248*, 2022. 2

[96] Zhentao Yu, Zixin Yin, Deyu Zhou, Duomin Wang, Finn Wong, and Baoyuan Wang. Talking head generation with probabilistic audio-to-visual diffusion priors. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7645–7655, 2023. 2

[97] Chenxu Zhang, Saifeng Ni, Zhipeng Fan, Hongbo Li, Ming Zeng, Madhukar Budagavi, and Xiaohu Guo. 3d talking face with personalized pose dynamics. *IEEE Transactions on Visualization and Computer Graphics*, 2021. 2

[98] Chenxu Zhang, Yifan Zhao, Yifei Huang, Ming Zeng, Saifeng Ni, Madhukar Budagavi, and Xiaohu Guo. Facial: Synthesizing dynamic talking face with implicit attribute learning. In *ICCV*, pages 3867–3876, 2021. 2

[99] Shiwei Zhang, Jiayu Wang, Yingya Zhang, Kang Zhao, Hangjie Yuan, Zhiwu Qin, Xiang Wang, Deli Zhao, and Jingren Zhou. I2vgen-xl: High-quality image-to-video synthesis via cascaded diffusion models. *arXiv preprint arXiv:2311.04145*, 2023. 2

[100] 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. 2, 5, 8

[101] Zhimeng Zhang, Lincheng Li, Yu Ding, and Changjie Fan. Flow-guided one-shot talking face generation with a high-resolution audio-visual dataset. In *CVPR*, pages 3661–3670, 2021. 2, 5

[102] Zhimeng Zhang, Zhipeng Hu, Wenjin Deng, Changjie Fan, Tangjie Lv, and Yu Ding. Dinet: Deformation inpainting network for realistic face visually dubbing on high resolution video. *arXiv preprint arXiv:2303.03988*, 2023. 2

[103] Hang Zhou, Yu Liu, Ziwei Liu, Ping Luo, and Xiaogang Wang. Talking face generation by adversarially disentangled audio-visual representation. In *AAAI*, pages 9299–9306, 2019. 2

[104] 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 *CVPR*, pages 4176–4186, 2021. 2, 5, 8

[105] 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. 2, 5, 8

DreamTalk: When Expressive Talking Head Generation Meets Diffusion Probabilistic Models

Supplementary Material

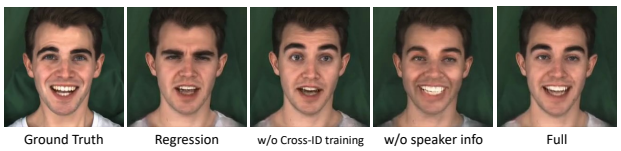

Figure 12. The qualitative results of style predictor's ablation study.

Method	SCD↓	MD↓	SA↑
w/o speaker info	0.49	0.28	64.3
w/o cross-ID training	0.68	0.45	28.1
regression	0.56	0.32	55.1
Full	**0.42**	**0.23**	**78.6**

Table 4. The ablation study results of the style predictor.

A. Additional Evaluation for Speaking Style Prediction

A.1. Ablation Study

To evaluate the impact of our design choices, we conduct an ablation study with three variants: (1) omitting speaker information and relying solely on audio for prediction (**w/o speaker info**); (2) during model training, the speaker info and audio are both obtained from the same video (**w/o cross-ID training**); (3) employing a regression model instead of a diffusion model for prediction (**regression**). Our full model is denoted as **Full**. When generating samples for evaluation, the facial images and audio we use are sourced from videos of the same individual expressing different emotions(*e.g.* the face image is from a happy video while the audio is from an angry one.). This generation approach better aligns with real-world applications.

How to quantitatively evaluate the performance of speaking style prediction has not been explored before. we devise three metrics:

- **Style Code Distance (SCD)** We extract the style codes from the videos that provide the audio input and compute the L2 distance between the predicted style codes and these style codes.
- **Motion Distance (MD)** We use the predicted style codes and the audio used for prediction to generate face motions and compute the L2 distance between the generated face motions and the face motions extracted from the ground truth videos.
- **Style Accuracy (SA)** We split test videos into different speaking styles and train a style classifier to classify which style a face motion sequence belongs to. Then, We classify the face motions generated using predicted style codes and report the accuracy. Specifically, we put the videos from the same speaker, emotion, and intensity into one style. We evaluate this metric on MEAD only since the number of RAVEDESS videos for each style is inadequate to train a style classifier. The ground truth testing set gets 92.5% accuracy.

We refrain from devising image-level metrics, such as training an image classifier for speaking style classification, due to several critical considerations. Firstly, factors in images that are irrelevant to expression, such as the speaker's identity and background elements, can adversely impact the accurate prediction of nuanced speaking styles. Secondly, inaccuracies introduced by the rendering process may further additionally hinder the accurate discernment of these subtle speaking styles.

The results are shown in Tab. 4 and Fig. 12. The **w/o speaker info** variant successfully predicts emotions from audio but occasionally fails to maintain consistency between the predicted speaking style and speaker identity, leading to poor identity preservation. This underscores the importance of speaker information in predicting speaking styles. Although in experiments, we observed that **w/o cross-ID training** achieves slightly better performance than **Full** when the input portrait and audio are from the same video, it underperformed, often failing to predict the correct emotion, when inputs were from different videos. This suggests that identity 3DMM parameters may convey some expression information, and without cross-ID training, the model might derive emotional cues from this leaked information rather than the audio. The **regression** variant struggles to generate accurate expressions for certain data, highlighting the superior distribution-learning capability of diffusion models in facilitating speaking style prediction.

A.2. User Study

In our user study, we evaluate the alignment between the original and predicted speaking styles. Directly assessing the alignment of speaking styles can be somewhat ambiguous, so we employ a comparative approach for evaluation. Specifically, we create a series of video triplets. Each triplet consisted of a test video from our dataset and two generated videos. The first video was generated using a style code predicted from an input portrait, sharing the same speaker identity as in the test video but displaying a neutral emotion, combined with the audio from the test video. The sec-

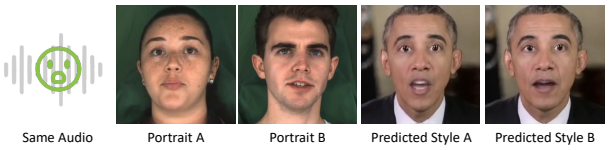

Figure 13. Analyzing the influence of portraits on style prediction. The audio conveys surprised emotion.

ond video is generated using the style code extracted from videos with the same emotion but from a speaker different from the one in the test video. We recruit 20 participants. Each participant is then asked to evaluate 20 triplets and identify which of the generated videos most accurately reflected the speaking style of the test video. The videos generated using predicted style codes are preferred in 75.8% of all ratings. This indicates that the style predictor is able to infer personalized speaking styles that are aligned with the audio.

A.3. Analyzing the Influence of Portraits

We analyze the influence of portraits on speaking style prediction by predicting speaking styles with an audio clip and different input portraits. The predicted styles are subsequently applied to an identical portrait for a clearer comparison. As shown in Fig. 13, the predicted speaking styles match the subtle identity characteristics, such as gender, of the input portraits. The predicted style A generated more feminine results. This validates the necessity of integrating portrait information during style prediction.

B. Additional Results for Expressive Talking Head Generation

B.1. Analysis on Generalization Capabilities

Songs. As demonstrated in *Supplementary Video*, our method successfully generates reasonable results for songs, even those with accompaniment, despite this being significantly different from the training dataset's data distribution. A noticeable decline in lip-sync accuracy is observed when the accompaniment volume is excessively high. We conduct a comparative analysis of lip-sync performance between songs with accompaniment and songs with removed accompaniment (using data from the SingFace Dataset [38]). It is found that the accompaniment adversely affects lip-sync, leading to mouth movements resembling mumbling. Addressing the negative impact of accompaniment on lip-sync accuracy presents an interesting avenue for future research.
Speech in Multiple Languages. *Supplementary Video* shows that our method generates satisfactory results with speech in French, Chinese, Spanish, German, Italian, Japanese, and Korean. The versatility of wav2vec features aids application across various languages. Additionally, the inclusion of multilingual talking head videos from Voxceleb2 enhances generalization.

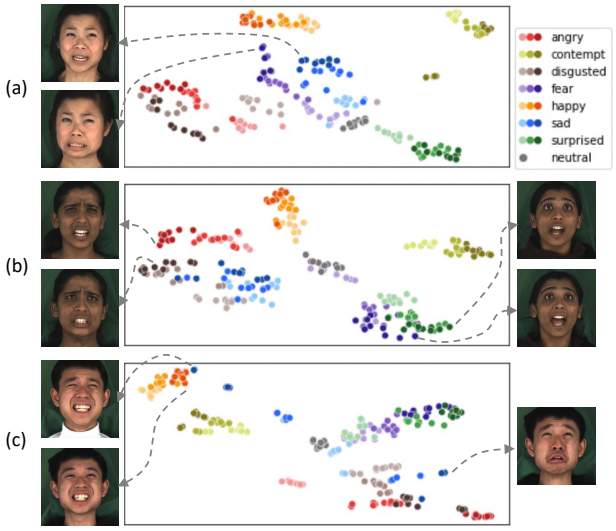

Figure 14. t-SNE visualization of style codes for 3 speakers, with darker hues representing increased emotional intensity.

Noisy Audio. *Supplementary Video* demonstrates that our method yields satisfactory outcomes when processing audio mixed with multiple noise types and intensities. We employ noise recordings from typical talking head application environments—meetings, offices, and cafeterias—sourced from the DEMAND dataset [73]. Using a publicly available tool[1], we blend the audio with noise at SNRs of 20 dB, 10 dB, and 0 dB. Remarkably, our method maintains performance even at 0 dB SNR, where the noise is as loud as the speech and significantly impairs speech intelligibility and clarity.

B.2. More Results of Style Code Visualization

We observe that each speaker's style code distribution exhibits both common patterns and individualized characteristics. Common patterns include: Firstly, speaking styles of different emotions cluster together first, with styles of lower intensity being closer to neutral and those of higher intensity being further away. Secondly, speaking styles of anger and disgust, as well as fear and surprise, often cluster together, as shown in Fig. 14 (b) and (c). Note that unlike Ma et al. [46], our method does not incorporate losses to constrain style space.

Fig. 14 (a) illustrates an example of individualized characteristics. The speaker's manifestation of fear closely resembles sadness, lacking the characteristic wide-eyed and open-mouthed expression, thereby positioning the speaking styles of fear nearer to those of sadness rather than surprise. Even within the same emotion, a speaker's speaking style

[1] https://github.com/Sato-Kunihiko/audio-SNR

can exhibit notable variation. In Fig. 14 (c), the speaker's dual expression of sadness—once with clenched teeth, similar to happy expressions, and another with depressed lip corners, akin to disgust—results in style codes close to the respective emotions. This observation diminishes the rationale for manually categorizing styles based on emotion and intensities in Ma et al. [46].

C. Implementation Details

C.1. Architectual Details

Denoising Network. The audio encoder processes an input window of 11 sequential audio features, each of dimension 1024. These features undergo dimension reduction to 256 via a linear layer and then are fed to a transformer encoder comprising three 8-head transformer encoder layers, each with a hidden size of 256. Subsequently, a linear layer transforms the output tokens to yield audio tokens sized 11×256. The audio tokens are concatenated with noisy motion and then added with the encoded diffusion step.

The style encoder ingests sequential expression parameters from style reference videos, each sequence sized $N \times 64$. These sequences, ranging in length from 64 to 256 frames, are initially expanded to 256 dimensions via a linear layer. Subsequently, they are introduced into a transformer encoder, composed of three 8-head layers, each with a hidden size of 256. The resulting output tokens, each with a dimension of 256, are aggregated through self-attention pooling [56], yielding a style code of dimension 256.

Within the decoder, the style code is repeated 11 times, subsequently added with positional embedding to produce style tokens. These tokens, in conjunction with audio tokens, are processed by a transformer decoder, encompassing three 8-head layers, each with a hidden dimension of 256. Here, style tokens serve as the query, while audio tokens serve as both key and value. The middle output token is fed into a linear layer to predict facial motion.

Style-aware Lip Expert. The face mesh is obtained by adding the mean shape to the product of the expression parameters and expression bases. Fig. 15 shows the architecture of the audio embedder and the mouth embedder.

Style Predictor. The style predictor is implemented as a transformer encoder comprising six 8-head transformer encoder layers, each with a hidden size of 256. The input features are all linearly projected to 256.

C.2. Data Details

C.2.1 Datasets

MEAD. The dataset, an in-lab talking-face corpus, features 60 speakers articulating eight emotions at three different intensity levels. When dividing the MEAD dataset into training and test subsets, we adhere to previously established methodologies [30].

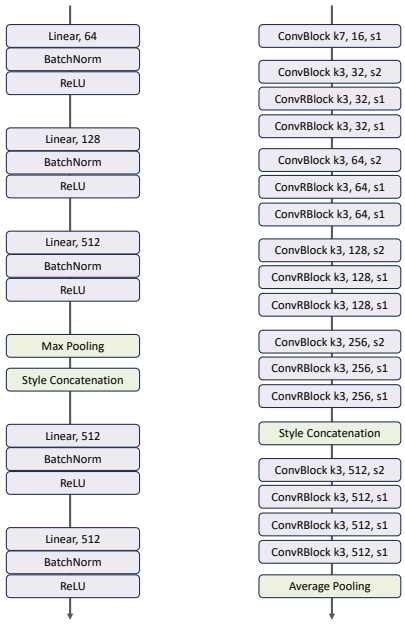

Figure 15. The architecture of mouth embedder and audio embedder. ConvBlock consists of a 1D CNN layer and a batch normalization layer. "k3, 32, s2" means that the kernel size is 3, the output dimension is 32, and the stride is 2. ConvRBlock is a ConvBlock with residual connection.

HDTF. The dataset stands out as a high-resolution, in-the-wild audio-visual dataset. We designate 10% of HDTF videos for testing and reserve the remainder for training.

Voxceleb2. Voxceleb2 is a large-scale talking head dataset with videos collected from YouTube. We redownload and recrop the videos to improve their resolution to 256×256. Subsequently, approximately 80000 high-quality videos are selected, with 400 allocated for testing and the rest for training.

RAVDESS. The dataset features 24 professional actors (12 female, 12 male) vocalizing two lexically matched statements in a neutral North American accent. It encompasses a range of expressions in both speech and song, each articulated at two levels of emotional intensity, with an additional neutral expression included. We employ the speech data from RAVDESS, encompassing eight emotions, for evaluation.

C.2.2 Data Processing

The original videos are cropped and resized to 256×256 pixels, aligning with the specifications in FOMM [59], and are sampled at 25 FPS. The 3DMM parameters are extracted by Deep3DFace [15].

Regarding the audio features used in the denoising network and the style-aware lip expert, we downsample the

speech wave into the sampling rate of 16000 and extract acoustic features employing a pre-trained Wav2Vec2.0 model [2].

For audio features used in the style predictor, we extract them using a pre-trained HuBERT model [28]. Besides, we also utilize low-level audio features including Mel Frequency Cepstrum Coefficients (MFCC), Mel-filterbank energy features (FBANK), fundamental frequency, and voice flag. These two type of features are concatenated to represent the audio features used in the style predictor.

C.3. Training Details

Our framework is implemented on Pytorch. We employ Adam [32] for optimization, with a learning rate set to 0.0001. The number of diffusion steps for the denoising network and style predictor is 1000. The training batch size for the denoising network, style predictor, and lip expert is 64, 64, and 32, respectively. λ_{denoise}, λ_{sync}, n, and w are set to 1, 1, 5, and 5, respectively. The number of frames for style reference and audio used in style prediction is limited to $64-256$, corresponding to a time length of $2.56-10.24$ seconds. The denoising network, style predictor, and lip expert are trained on one NVIDIA Tesla A100 GPU for about 3, 1, and 10.5 hours, respectively.

C.3.1 Finetuning PIRender

The renderer is fine-tuned with the losses in Ren et al. [52] using MEAD. Instead of training with the self-reconstruction protocol where the source frame and target frame are from the same video, we select the source frame and target frame from the same speaker with different emotions. This enables the renderer to generate emotions different from the input portrait. This also allows our method to utilize portraits with emotions, unlike previous approaches that are confined to using neutral portraits [30]. We also observed that when fine-tuned only on the emotional talking head dataset, the renderer struggles with identity preservation. We argue that this problem stems from the fact that the number of speakers in the emotional dataset is limited. Therefore, we incorporate some neutral videos in Voxceleb into the data used for fine-tuning. This enhances the performance in identity preservation.

C.3.2 Training Style-aware Lip Expert

The style-aware lip expert is trained to discriminate whether the input audio and face motions are synchronized. We use cosine-similarity with binary cross-entropy loss to train the lip expert. Specifically, we compute cosine-similarity for the face motion embedding $\boldsymbol{e}^m$ and audio embedding $\boldsymbol{e}^a$ to represent the probability that the input audio-motion pair is synchronized. The training loss of the lip expert is:

$$\mathcal{L}_{\text{expert}} = \text{BCE}(\frac{\boldsymbol{e}^m \cdot \boldsymbol{e}^a}{\max(||\boldsymbol{e}^m||_2 \cdot ||\boldsymbol{e}^a||_2, \epsilon)}), \tag{10}$$

where ϵ is a small number for avoiding the division-by-zero error.

C.4. Inference Details

The inference of the denoising network can be accelerated with DDIM. We generate samples with 10 DDIM steps and observe no performance drop. Generating a 30-second video offline takes 15.61 seconds, with the face motion generation only taking 1.24 seconds. During evaluation, the scale factor ω of classifier-free guidance is set to 1. The style predictor uses the sampling algorithm of DDPM to predict style codes.

The emotion conveyed in the style reference (video or audio), should remain consistent to avoid confusing the model.

The head pose information, which is fed into the renderer, can be derived from real videos or generated using existing methods [83].

D. Limitations and Future Work

Despite DreamTalk's promising advancements in expressive talking head generation, it encounters several challenges that open avenues for future research.

Firstly, the method occasionally produces artifacts, such as teeth flickering, around the mouth area, particularly during intense expressions. Generating teeth is a long-standing challenge in talking head generation since the algorithm needs to inpaint the teeth area that are often occluded in the input portrait. This problem is exacerbated under intense expressions where the teeth area expands. The issue can be mitigated by incorporating modules [36, 44, 92] proposed recently that enhance the teeth quality. A more comprehensive solution involves developing an emotion-specific renderer, as current renderers [20, 30, 46, 90] are mainly adaptations of existing facial reenactment methods with minimal modifications. An emotion-aware renderer would not only address the teeth generation issue but also enhance the overall expressiveness of emotions.

Secondly, DreamTalk does not account for the temporal variability in speaking styles. In real-life scenarios, a speaker's style evolves over time, a feature our method currently overlooks. For instance, at the end of a speech, our method might still produce expressions of intense emotion, such as a wide-open mouth in surprise, instead of a more neutral, closed-mouth expression. Introducing a module that dynamically predicts speaking style over time could address this limitation.

Thirdly, the style predictor sometimes struggles with accurately identifying emotions in low-intensity audio clips

from the MEAD dataset (some audio in MEAD intensity level 1 videos). The emotion conveyed in these audio clips is very similar to neutral emotion and thus confusing the prediction. We also observed that although the videos in the MEAD dataset clearly express emotions due to the good training and supervision of the speakers, sometimes the audio does not correspond with the emotions that should be expressed. Therefore, To enhance prediction accuracy, employing a dataset where the audio closely aligns with the expressed emotions could be beneficial. Another solution is to incorporate textual information from audio during prediction, a strategy commonly employed in speech emotion recognition [8, 85].

Despite these challenges, DreamTalk marks a significant stride in the realm of high-quality, expressive talking head generation, setting a foundation for further innovations.

E. Ethical Consideration

DreamTalk is able to generate realistic talking head videos. This positions DreamTalk with a broad spectrum of potential applications, each carrying intricate societal implications. While DreamTalk holds significant potential in amplifying and enriching human creative endeavors and may pave the way for innovative tools for creative professionals, its capabilities also harbor risks. There's a possibility for DreamTalk to generate content that might encompass or imply sexual themes, promote hatred, or depict violence. Misuse of DreamTalk could lead to negative repercussions on individuals or groups, potentially erasing or maligning them, perpetuating stereotypes, and subjecting them to disrespect. Other concerns include the potential for harassment, intimidation, or exploitation. Furthermore, DreamTalk's capabilities might be harnessed to mislead or spread misinformation.

Before releasing DreamTalk, we have implemented and plan to introduce several safeguards to curb potential misuse. We've purged detrimental content from the training dataset and incorporated visual filters to deter users from creating harmful outputs. To counteract biases in the generated results, we're enhancing the dataset's diversity by manually ensuring balance, which will reduce instances of erasure, stereotype perpetuation, indignity, and uneven quality across inputs. Users will be advised against using images without the depicted individuals' consent to combat harassment and bullying. To prevent the spread of misinformation, all DreamTalk outputs will bear watermarks indicating their synthetic nature. Our pre-release strategy involves a thorough risk assessment, leveraging a growing suite of safety evaluations and red teaming techniques. We'll also scrutinize the findings from pilot tests centered on new use cases and conduct in-depth post-release assessments. Both automated and manual monitoring mechanisms are in development to preempt misuse. Our commitment remains steadfast in continuously researching ways to minimize adverse societal effects.