EDTalk: Efficient Disentanglement for Emotional Talking Head Synthesis

Shuai Tan[1], Bin Ji[1], Mengxiao Bi[2], and Ye Pan[1]*

[1] Shanghai Jiao Tong University
[2] NetEase Fuxi AI Lab
{tanshuai0219, bin.ji, whitneypanye}@sjtu.edu.cn
bimengxiao@corp.netease.com

Abstract. Achieving disentangled control over multiple facial motions and accommodating diverse input modalities greatly enhances the application and entertainment of the talking head generation. This necessitates a deep exploration of the decoupling space for facial features, ensuring that they **a)** operate independently without mutual interference and **b)** can be preserved to share with different modal inputs—both aspects often neglected in existing methods. To address this gap, this paper proposes a novel **E**fficient **D**isentanglement framework for **Talk**ing head generation (**EDTalk**). Our framework enables individual manipulation of mouth shape, head pose, and emotional expression, conditioned on video or audio inputs. Specifically, we employ three **lightweight** modules to decompose the facial dynamics into three distinct latent spaces representing mouth, pose, and expression, respectively. Each space is characterized by a set of learnable bases whose linear combinations define specific motions. To ensure independence and accelerate training, we enforce orthogonality among bases and devise an **efficient** training strategy to allocate motion responsibilities to each space without relying on external knowledge. The learned bases are then stored in corresponding banks, enabling shared visual priors with audio input. Furthermore, considering the properties of each space, we propose an Audio-to-Motion module for audio-driven talking head synthesis. Experiments are conducted to demonstrate the effectiveness of EDTalk. The code and pretrained models are released at: https://tanshuai0219.github.io/EDTalk/

Keywords: Talking head generation · Facial disentanglement

1 Introduction

Talking head animation has garnered significant research attention owing to its wide-ranging applications in education, filmmaking, virtual digital humans, and the entertainment industry [35]. While previous methods [23, 39, 56, 57] have achieved notable advancements, most of them generate talking head videos in a holistic manner, lacking fine-grained individual control. Consequently, attaining

* Corresponding author.

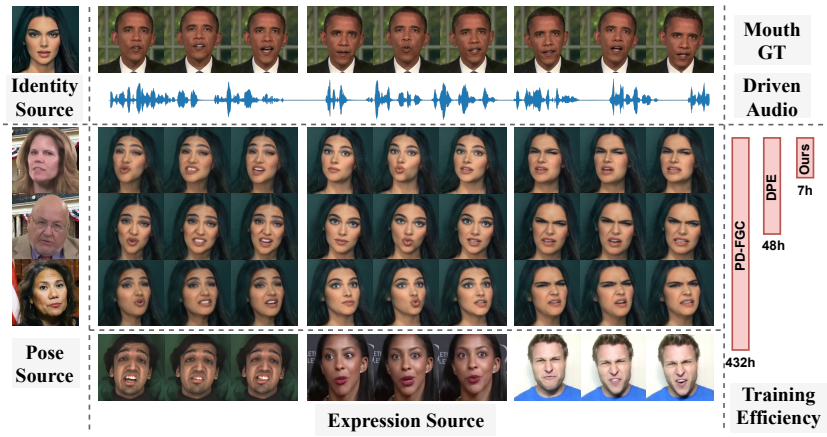

Fig. 1: Illustrative animations produced by EDTalk. Given an identity source, EDTalk synthesizes talking face videos characterized by mouth shapes, head poses, and expressions consistent with mouth GT, pose source and expression source. These facial dynamics can also be inferred directly from driven audio. Importantly, EDTalk demonstrates superior efficiency in disentanglement training compared to other methods.

precise and disentangled manipulation over various facial motions such as mouth shapes, head poses, and emotional expressions remains a challenge, crucial for crafting lifelike avatars [47]. Moreover, existing approaches typically cater to only one driving source: either audio [28, 46] or video [17, 41], thereby limiting their applicability in the multimodal context. There is a pressing need for a unified framework capable of simultaneously achieving individual facial control and handling both audio-driven and video-driven talking face generation.

To tackle the challenges, an intuition is to disentangle the entirety of facial dynamics into distinct facial latent spaces dedicated to individual components. However, it is non-trivial due to the intricate interplay among facial movements [47]. For instance, mouth shapes profoundly impact emotional expressions, where one speaks happily with upper lip corners but sadly with the depressed ones [13, 14]. Despite the extensive efforts made in facial disentanglement by previous studies [27, 33, 47, 58, 65], we argue there exist three key limitations. **(1)** Overreliance on external and prior information increases the demand for data and complicates the data pre-processing: One popular line [27, 47, 58] relies heavily on external audio data to decouple the mouth space via contrastive learning [24]. Subsequently, they further disentangle the pose space using predefined 6D pose coefficients extracted from 3D face reconstruction models [11]. However, such external and prior information escalates dataset demands and any inaccuracies therein can lead to the trained model errors. **(2)** Disentangling latent spaces without internal constraints leads to incomplete decoupling. Previous works [27, 65] simply constrain each space externally with a prior during the decoupling process, overlooking inter-space constraints. This oversight fails

to ensure that each space exclusively handles its designated component without interference from others, leading to training complexities, reduced efficiency, and performance degradation. **(3)** Inefficient training strategy escalates the training time and computational cost. When disentangling a new sub-space, some methods [33, 47] require training the entire heavyweight network from scratch, which significantly incurs high time and computational costs [14]. It can be costly and unaffordable for many researchers. Furthermore, most methods are unable to utilize audio and video inputs simultaneously.

To cope with such issues, this paper proposes an **E**fficient **D**isentanglement framework, tailored for one-shot talking head generation with precise control over mouth shape, head pose, and emotional expression, conditioned on video or audio inputs. Our key insight lies in our requirements for decoupled space: **(a)** The decoupled spaces should be disjoint, which means each space captures solely the motion of its corresponding component without the interference from others. This also ensures that decoupling a new space will not affect the trained models, thereby avoiding the necessity of training from scratch. **(b)** Once the spaces are disentangled from video data to support video-driven paradigm, they should be stored to share with the audio inputs for further audio-driven setting.

To this end, drawing inspiration from the observation that the entire motion space can be represented by a set of directions [53], we innovatively disentangle the whole motion space into three distinct component-aware latent spaces. Each space is characterized by a set of learnable bases. To ensure that different latent spaces do not interfere with each other, we constrain bases orthogonal to each other not only *intra*-space [53] but also *inter*-space. To accomplish the disentanglement without prior information, we introduce a progressive training strategy comprising cross-reconstruction mouth-pose disentanglement and self-reconstruction complementary learning for expression decoupling. Despite comprising two stages, our decoupling process involves training only the proposed lightweight Latent Navigation modules, keeping the weights of other heavier modules fixed for efficient training.

To explicitly preserve the disentangled latent spaces, we store the base sets of disentangled spaces in the corresponding banks. These banks serve as repositories of prior bases essential for audio-driven talking head generation. Consequently, we introduce an Audio-to-Motion module designed to predict the weights of the mouth, pose, and expression banks, respectively. Specifically, we employ an audio encoder to synchronize lip motions with the audio input. Given the non-deterministic nature of head motions [61], we utilize normalizing flows [37] to generate probabilistic and realistic poses by sampling from a Gaussian distribution, guided by the rhythm of audio. Regarding expression, we aim to extract emotional cues from the audio [21] and transcripts. It ensures that the generated talking head video aligns with the tone and context of audio, eliminating the need for additional expression references. In this way, our EDTalk enables talking face generation directly from the sole audio input.

Our contributions are outlined as follows: **1)** We present EDTalk, an efficient disentanglement framework enabling precise control over talking head synthesis

concerning mouth shape, head pose, and emotional expression. **2)** By introducing orthogonal bases and an efficient training strategy, we successfully achieve complete decoupling of these three spaces. Leveraging the properties of each space, we implement Audio-to-Motion modules to facilitate audio-driven talking face generation. **3)** Extensive experiments demonstrate that our EDTalk surpasses the competing methods in both quantitative and qualitative evaluation.

2 Related Work

2.1 Disentanglement on the face

Facial dynamics typically involve coordinated movements such as head poses, mouth shapes, and emotional expressions in a global manner [45], making their separate control challenging. Several works have been developed to address this issue. PC-AVS [65] employs contrastive learning to isolate the mouth space related to audio. Yet since similar pronunciations tend to correspond to the same mouth shape [26], the constructed negative pairs in a mini-batch often include positive pairs and the number of negative pairs in the mini-batch is too small [16], both of which results in subpar results. Similarly, PD-FGC [47] and TH-PAD [58] face analogous challenges in obtaining content-related mouth spaces. Although TH-PAD incorporates lip motion decorrelation loss to extract non-lip space, it still retains a coupled space where expressions and head poses are intertwined. This coupling results in randomly generated expressions co-occurring with head poses, compromising user-friendliness and content relevance. Despite the achievement of PD-FGC in decoupling facial details, its laborious coarse-to-fine disentanglement process consumes substantial computational resources and time. DPE [33] introduces a bidirectional cyclic training strategy to disentangle head pose and expression from talking head videos. However, it necessitates two generators to independently edit expression and pose sequentially, escalating computational resource consumption and runtime. In contrast, we propose an efficient decoupling approach to segregate faces into mouth, head pose, and expression components, readily controllable by different sources. Moreover, our method requires only a unified generator, and minimal additional resources are needed when exploring a new disentangled space.

2.2 Audio-driven Talking Head Generation

Audio-driven talking head generation [3, 29] endeavors to animate images with accurate lip movements synchronized with input audio clips. Research in this area is predominantly categorized into two groups: intermediate representation based methods and reconstruction-based methods. Intermediate representation based methods [4, 6, 12, 51, 52, 56, 59, 63, 66] typically consist of two submodules: one predicts intermediate representations from audio, and the other synthesizes photorealistic images from these representations. For instance, Das et al. [12] employ landmarks as an intermediate representation, utilizing an audio-to-landmark module and a landmark-to-image module to connect audio inputs

and video outputs. Yin et al. [57] extract 3DMM parameters [2] to warp source images using predicted flow fields. However, obtaining such intermediate representations, like landmarks and 3D models, is laborious and time-consuming. Moreover, they often offer limited facial dynamics details, and training the two sub-modules separately can accumulate errors, leading to suboptimal performance. In contrast, our approach operates within a reconstruction-based framework [5, 7, 39, 40, 44, 46, 49, 64]. It integrates features extracted by encoders from various modalities to reconstruct talking head videos in an end-to-end manner, alleviating the aforementioned issues. A notable example is Wav2Lip [36], which employs an audio encoder, an identity encoder, and an image decoder to generate precise lip movements. Similarly, Zhou et al. [65] incorporate an additional pose encoder for free pose control, yet disregard the nondeterministic nature of natural movement. To address this, we propose employing a probabilistic model to establish a distribution of non-verbal head motions. Additionally, none of the existing methods consider facial expressions, crucial for authentic talking head generation. Our approach aims to integrate facial expressions into the model to enhance the realism and authenticity of the generated talking heads.

2.3 Emotional Talking Head Generation

Emotional talking head generation is gaining traction due to its wide-ranging applications and heightened entertainment potential. On the one hand, some studies [14,21,43,50] identify emotions using discrete emotion labels, albeit facing a challenge to generate controllable and fine-grained expressions. On the other hand, recent methodologies [20,27,31,47] incorporate emotional images or videos as references to indicate desired expressions. Ji et al. [20], for instance, mask the mouth region of an emotional video and utilize the remaining upper face as an expression reference for emotional talking face generation. However, as mouth shape plays a crucial role in conveying emotion [45], they struggle to synthesize vivid expressions due to their failure to decouple expressions from the entire face. Thanks to our orthogonal base and efficient training strategy, we are capable of fully disentangling different motion spaces like mouth shape and emotional expression, thus achieving finely controlled talking head synthesis. Moreover, we also incorporate emotion contained within audio and transcripts. To the best of our knowledge, we are the first to achieve this goal—automatically inferring suitable expressions from audio tone and text, thereby generating consistent emotional talking face videos without relying on explicit image/video references.

3 Methodology

As illustrated in Fig. 2 (a), given an identity image I^i, we aim to synthesize emotional talking face image $\hat{I}^g$ that maintains consistency in identity information, mouth shape, head pose, and emotional expression with various driving sources I^i, I^m, I^p and I^e. Our intuition is to disentangle different facial components from the overall facial dynamics. To this end, we propose EDTalk (Sec. 3.1)

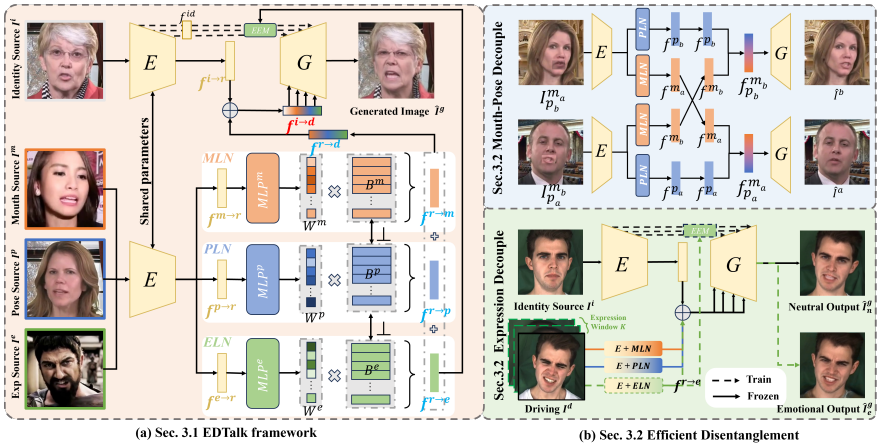

Fig. 2: Illustration of our proposed EDTalk. (a) EDTalk framework. Given an identity source I^i and various driving images I^* ($* \in \{m, p, e\}$) for controlling corresponding facial components, EDTalk animates the identity image I^i to mimic the mouth shape, head pose, and expression of I^m, I^p and I^e with the assistance of three Component-aware Latent Navigation modules: MLN, PLN and ELN. (b) Efficient Disentanglement. The disentanglement process consists of two parts: Mouth-Pose decouple and Expression Decouple. For the former, we introduce the cross-reconstruction training strategy aimed at separating mouth shape and head pose. For the latter, we achieve expression disentanglement using self-reconstruction complementary learning.

with learnable orthogonal bases stored in banks B^* ($*$ refers to the mouth source m, pose source p and expression source e for simplicity), each representing a distinct direction of facial movements. To ensure the bases are component-aware, we propose an efficient disentanglement strategy (Sec. 3.2), comprising Mouth-Pose Decoupling and Expression Decoupling, which decompose the overall facial motion into mouth, pose, and expression spaces. Leveraging these disentangled spaces, we further explore an Audio-to-Motion module (Section 3.3, Figure 3) to produce audio-driven emotional talking face videos featuring probabilistic poses, audio-synchronized lip motions, and semantically-aware expressions.

3.1 EDTalk Framework

Figure 2 (a) illustrates the structure of EDTalk, which is based on an autoencoder architecture consisting of an Encoder E, three Component-aware Latent Navigation modules (CLNs) and a Generator G. The encoder E maps the identity image I^i and various driving source I^* into the latent features $f^{i\to r} = E(I^i)$ and $f^{*\to r} = E(I^*)$. The process in inspired by FOMM [41] and LIA [53]. Instead of directly modeling motion transformation $f^{i\to *}$ from identity image I^i to driving image I^* in the latent space, we posit the existence of a canonical feature f^r, that facilitates motion transfer between identity features and driving ones, expressed as $f^{i\to *} = f^{i\to r} + f^{r\to *}$.

Thus, upon acquiring the latent features $f^{*\to r}$ extracted by E from driving images I^*, we devise three Component-aware Latent Navigation modules to transform them into $f^{r\to *} = CLN(f^{*\to r})$. For clarity, we use pose as an example, denoted as $* = p$. Within the Pose-aware Latent Navigation (PLN) module, we establish a pose bank $B^p = \{b_1^p, ..., b_n^p\}$ to store n learnable base b_i^p. To ensure each base represents a distinct pose motion direction, we enforce orthogonality between every pair of bases by imposing a constraint of $\langle b_i^p, b_j^p \rangle = 0 \quad (i \neq j)$, where $\langle \cdot, \cdot \rangle$ signifies the dot product operation. It allows us to depict various head pose movements as linear combinations of the bases. Consequently, we design a Multi-Layer Perceptron layer MLP^p to predict the weights $W^p = \{w_1^p, ..., w_n^p\}$ of the pose bases from the latent feature $f^{p\to r}$:

$$W^p = \{w_1^p, ..., w_n^p\} = MLP^p(f^{p\to r}), \qquad f^{r\to p} = \sum_{i=1}^{n} w_i^p b_i^p, \tag{1}$$

Mouth and Expression-aware Latent Navigation module share the same architecture with PLM but have different parameters, where we can also derive $f^{r\to m} = \sum_{i=1}^{n} w_i^m b_i^m, W^m = MLP^m(f^{m\to r})$ and $f^{r\to e} = \sum_{i=1}^{n} w_i^e b_i^e, W^e = MLP^e(f^{e\to r})$ in the similar manner. It's worth noting that to achieve complete disentanglement of facial components and prevent changes in one component from affecting others, we ensure orthogonality between the three banks (B^m, B^p, B^e). This also allows us to directly combine the three features to obtain the driving feature $f^{r\to d} = f^{r\to m} + f^{r\to p} + f^{r\to e}$. We further get $f^{i\to d} = f^{i\to r} + f^{r\to d}$, which is subsequently fed into the Generator G to synthesize the final result $\hat{I}^g$. To maintain identity information, G incorporates the identity features f^{id} of the identity image via skip connections. Additionally, to enhance emotional expressiveness with the assistance of the emotion feature $f^{r\to e}$, we introduce a lightweight plug-and-play Emotion Enhancement Module (EEM), which will be discussed in the subsequent subsection. In summary, the generation process can be formulated as follows:

$$\hat{I}^g = G(f^{i\to d}, f^{id}, EEM(f^{r\to e})), \tag{2}$$

where EEM is exclusively utilized during emotional talking face generation. For brevity, we omit f^{id} in the subsequent equations.

3.2 Efficient Disentanglement

Based on the outlined framework, the crux lies in training each Component-aware Latent Navigation module to store only the bases corresponding to the motion of its respective components and to ensure no interference between different components. To achieve this, we propose an efficient disentanglement strategy comprising Mouth-Pose Decoupling and Expression Decoupling, thereby separating the overall facial dynamics into mouth, pose, and expression components.
Mouth-Pose Decouple. As depicted at the top of Fig. 2 (b), we introduce cross-reconstruction technical, which involves synthesized images of switched

mouths: $I_{p_b}^{m_a}$ and $I_{p_a}^{m_b}$. Here, we superimpose the mouth region of I^a onto I^b and vice versa. Subsequently, the encoder E encodes them into canonical features, which are processed through PLN and MLN to obtain corresponding features:

$$f^{p_b}, f^{m_a} = PLN(E(I_{p_b}^{m_a})), MLN(E(I_{p_b}^{m_a})) \tag{3}$$

$$f^{p_a}, f^{m_b} = PLN(E(I_{p_a}^{m_b})), MLN(E(I_{p_a}^{m_b})) \tag{4}$$

Next, we substitute the extracted mouth features and feed them into the generator G to perform cross reconstruction of the original images: $\hat{I}^b = G(f^{p_b}, f^{m_b})$ and $\hat{I}^a = G(f^{p_a}, f^{m_a})$. Additionally, we include identity features f^{id} extracted from another frame of the same identity as input to the generator G. Afterward, we supervise the Mouth-Pose Decouple module by adopting reconstruction loss $\mathcal{L}_{\text{rec}}$, perceptual loss $\mathcal{L}_{\text{per}}$ [22, 60] and adversarial loss $\mathcal{L}_{\text{adv}}$:

$$\mathcal{L}_{\text{rec}} = \sum_{\#=a,b} \|I^{\#} - \hat{I}^{\#}\|_1; \qquad \mathcal{L}_{\text{per}} = \sum_{\#=a,b} \|\Phi(I^{\#}) - \Phi(\hat{I}^{\#})\|_2^2; \tag{5}$$

$$\mathcal{L}_{\text{adv}} = \sum_{\#=a,b} (\log D(I^{\#}) + \log(1 - D(\hat{I}^{\#}))), \tag{6}$$

where Φ denotes the feature extractor of VGG19 [42] and D is a discriminator tasked with distinguishing between reconstructed images and ground truth (GT). In addition, self-reconstruction of the Ground Truth (GT) is crucial, where mouth features and pose features are extracted from the same image and then input into G to reconstruct itself using $\mathcal{L}_{\text{self}}$. Furthermore, we impose feature-level constraints on the network:

$$\mathcal{L}_{\text{fea}} = \sum_{\#=a,b} (exp(-\mathcal{S}(f^{p_{\#}}, PLN(E(I^{\#})))) + exp(-\mathcal{S}(f^{m_{\#}}, MLN(E(I^{\#}))))), \tag{7}$$

where we extract mouth features and pose features from I^a and I^b, aiming to minimize their disparity with those extracted from synthesized images of switched mouths using cosine similarity $\mathcal{S}(\cdot, \cdot)$. Once the losses have converged, the parameters are no longer updated for the remainder of training, significantly reducing training time and resource consumption for subsequent stages.

Expression Decouple. As illustrated in the bottom of Fig. 2 (b), to decouple expression information from driving image I^d, we introduce Expression-aware Latent Navigation module (ELN) and a lightweight plug-and-play Emotion Enhancement Module (EEM), both trained via self-reconstruction complementary learning. Specifically, given an identity source I^i and a driving image I^d sharing the same identity as I^i but differing in mouth shapes, head poses and emotional expressions, our pre-trained modules (i.e., E, MLN, PLN, and G) from previous stage effectively disentangle mouth shape and head pose from I^d and drive I^i to generate $\hat{I}_n^g$ with matching mouth shape and head pose as I^d but with the same expression with I^i. Therefore, to faithfully reconstruct I^d with the same

expression, ELN is compelled to learn complementary information not disentangled by MLN, PLN, precisely the expression information. Motivated by the observation [47] that expression variation in a video sequence is typically less frequent than changes in other motions, we define a window of size K around I^d and average K extracted expression features to obtain a clean expression feature $f^{r\to e}$. $f^{r\to e}$ is then combined with extracted mouth and pose features as input to the generator G. Additionally, EEM takes $f^{r\to e}$ as input and utilizes affine transformations to produce $f^e = (f_s^e, f_b^e)$ that control adaptive instance normalization (AdaIN) [19] operations. The AdaIN operations further adapt identity feature f^{id} as emotion-conditioned features f_e^{id} by:

$$f_e^{id} := EEM(f^{id}) = f_s^e \frac{f^{id} - \mu(f^{id})}{\sigma(f^{id})} + f_b^e, \tag{8}$$

where $\mu(\cdot)$ and $\sigma(\cdot)$ represent the average and variance operations. Subsequently, we generate output $\hat{I}_e^g$ with the expression of I^d via Eq. 2. We enforce a motion reconstruction loss [47] $\mathcal{L}_{\text{mot}}$ in addition to the same reconstruction loss $\mathcal{L}_{\text{rec}}$, perceptual loss $\mathcal{L}_{\text{per}}$ and adversarial loss $\mathcal{L}_{\text{adv}}$ as Eq. 5 and Eq. 6:

$$\mathcal{L}_{\text{mot}} = \|\phi(I^d) - \phi(\hat{I}_e^g)\|_2 + \|\psi(I^d) - \psi(\hat{I}_e^g)\|_2, \tag{9}$$

where $\phi(\cdot)$ and $\psi(\cdot)$ denote features extracted by the 3D face reconstruction network and the emotion network of [11]. Moreover, to ensure that the synthesized image accurately mimics the mouth shape of the driving frame, we further introduce a mouth consistency loss $\mathcal{L}_{\text{m-c}}$:

$$\mathcal{L}_{\text{m-c}} = e^{-\mathcal{S}(MLN(E(\hat{I}_e^g)), MLN(E(I^d)))}, \tag{10}$$

where MLN and E are pretrained in the previous stage. During training, we only need to train lightweight ENL and EEM, resulting in fast training.

After successfully training the two-stage Efficient Disentanglement module, we acquire three disentangled spaces, enabling one-shot video-driven talking face generation with separate control of identity, mouth shape, pose, and expression, given different driving sources, as illustrated in Fig. 2 (a).

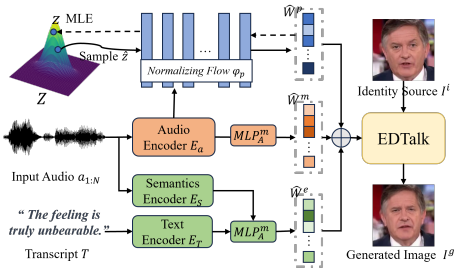

Fig. 3: The overview of Audio-to-Motion, for mouth, pose, expression prediction.

3.3 Audio-to-Motion

Integrating the disentangled spaces, we aim to address a more appealing but challenging task: audio-driven talking face generation. In this section, depicted in Fig. 3, we introduce

three modules to predict the weights of pose, mouth, and expression from audio. These modules replace the driving video input, facilitating audio-driven talking face generation.

Audio-Driven Lip Generation. Prior works [31,45] generate facial dynamics, encompassing lip motions and expressions, in a holistic manner, which proves challenging for two main reasons: 1) Expressions, being acoustic-irrelevant motions, can impede lip synchronization [61]. 2) The absence of lip visual information hinders fine details synthesis at the phoneme level [34]. Thanks to the disentangled mouth space obtained in the previous stage, we naturally mitigate the influence of expression without necessitating special training strategies or loss functions like [61]. Additionally, since the decoupled space is trained during video-driven talking face generation using video as input, which offers ample visual information in the form of mouth bases b_i^m stored in the bank B^m, we eliminate the need for extra visual memory like [34]. Instead, we only need to predict the weight w_i^m of each base b_i^m, which generates the fine-grained lip motion. To achieve this, we design an Audio Encoder E_a, which embeds the audio feature into a latent space $f^a = E_a(a_{1:N})$. Subsequently, a linear layer MLP_A^m is added to decode the mouth weight $\hat{W}^m$. During training, we fix the weights of all modules and only update E_a and MLP_A^m using the weighted sum of feature loss $\mathcal{L}_{fea}^m$, reconstruction loss $\mathcal{L}_{rec}^m$ and sync loss $\mathcal{L}_{sync}^m$ [36]:

$$\mathcal{L}_{fea}^m = \|W^m - \hat{W}^m\|_2, \qquad \mathcal{L}_{rec}^m = \|I - \hat{I}\|_2, \tag{11}$$

$$\mathcal{L}_{sync}^m = -\log(\frac{v \cdot s}{max(\|v\|_2 \cdot \|s\|_2, \epsilon)}), \tag{12}$$

where $W^m = MLN(E(I))$ is the GT mouth weight extracted from GT image I and $\hat{I}$ is generated image using Eq. 2. $\mathcal{L}_{sync}^m$ is introduced from [36], where v and s are extracted by the speech encoder and image encoder in SyncNet [10].

Flow-Based Probabilistic Pose Generation. Due to the nature of one-to-many mapping from the input audio to head poses, learning a deterministic mapping like previous works [51, 52, 66] output the same results, which bring ambiguity and inferior visual results. To generate probabilistic and realistic head motions, we predict the pose weights $\hat{W}^p$ using Normalizing Flow φ_p [37], as illustrated in Fig. 3. During training (indicated by dash lines), we extract pose weights W^p from videos as the ground truth and feed them into our φ_p. By incorporating Maximum Likelihood Estimation (MLE) in Eq. 13, we embed it into a Gaussian distribution p_Z conditioned on audio feature $f^a = E_a(a_{1:N})$:

$$z_t = \varphi_p^{-1}(w_t^p, f_t^a), \qquad \mathcal{L}_{\text{MLE}} = -\sum_{t=0}^{N-1} \log p_{\mathcal{Z}}(z_t) \tag{13}$$

As the normalizing flow φ_p is bijective, we reconstruct the pose weight $\hat{W}^p = \varphi_p(z, f_t^a)$ and utilize a pose reconstruction loss $\mathcal{L}_{\text{rec}}^p$ along with a temporal loss

$\mathcal{L}^p_{\text{tem}}$ to constrain φ_p:

$$\mathcal{L}^p_{\text{rec}} = \|W^p - \hat{W}^p\|_2, \qquad \mathcal{L}_{\text{tem}} = \frac{1}{N-1}\sum_{t=1}^{N-1} \|(w^p_t - w^p_{t-1}) - (\hat{w}^p_t - \hat{w}^p_{t-1})\|_2 \quad (14)$$

During inference, we randomly sample $\hat{z}$ from the constructed distribution p_Z and then generate pose weights $\hat{W}^p = \varphi_p(z, f^a_t)$. This process ensures the diversity of head motions while maintaining consistency with the audio rhythm.

Semantically-Aware Expression Generation. As finding videos with a desired expression may not always be feasible, potentially limiting their application [30], we aim to explore the emotion contained in audio and transcript with the aid of the introduced Semantics Encoder E_S and Text Encoder E_T. Inspired by [54], our Semantics Encoder E_S is constructed upon the pretrained HuBERT model [18], which consists of a CNN-based feature encoder and a transformer-based encoder. We freeze the CNN-based feature encoder and only fine-tuned the transformer blocks. Text Encoder E_T is inherited from the pretrained Emoberta [25], which encodes the overarching emotional context embedded within textual descriptions. We concatenate the embeddings generated by E_S and E_T and feed them into a MLP^e_A to generate the expression weights $\hat{W}^e$. Since audio or text may not inherently contain emotion during inference, such as in TTS-generated speech, in order to support the prediction of emotion from a single modality, we randomly mask ($\mathcal{M}$) a modality with probability p during training, inspired by HuBERT:

$$\hat{W}^e = \begin{cases} MLP^e_a(E_S(a), E_T(T)), & 0.5 \le p \le 1, \\ MLP^e_a(\mathcal{M}(E_S(a)), E_T(T)), & 0.25 \le p < 0.5, \\ MLP^e_a(E_S(a), \mathcal{M}(E_T(T))), & 0 \le p < 0.25. \end{cases} \quad (15)$$

We employ $\mathcal{L}_{\text{exp}} = \|W^e - \hat{W}^e\|_1$ to encourage $\hat{W}^e$ close to weight W^e generated by pretrained ELN from emotional frames. Until now, we are able to generate **probabilistic semantically-aware** talking head videos solely from an identity image and the driving audio.

4 Experiments

4.1 Experimental Settings

Implement Details. Our model is trained and evaluated on the datasets MEAD [50] and HDTF [62]. Additionally, we report results on additional datasets, including LRW [9] and Voxceleb2 [8], for further assessment of our method in the supplementary. All video frames are cropped following FOMM [41] and resized to 256×256. Our method is implemented using PyTorch and trained using the Adam optimizer on 2 NVIDIA GeForce GTX 3090 GPUs. The dimension of the latent code $f^{*\to r}$ and bases b^* is set to 512, and the number of bases of B^m, B^p and B^e are set to 20, 6 and 10, respectively. The weight for $\mathcal{L}_{\text{mot}}$ is set to 10 and the remaining weights are set to 1.

Method	MEAD [50]						HDTF [62]				
	PSNR↑	SSIM↑	M/F-LMD↓	FID↓	$\text{Sync}_{\text{conf}}$ ↑	Acc_{emo} ↑	PSNR↑	SSIM↑	M/F-LMD↓	FID↓	$\text{Sync}_{\text{conf}}$ ↑
MakeItTalk [66]	19.442	0.614	2.541/2.309	37.917	5.176	14.64	21.985	0.709	2.395/2.182	18.730	4.753
Wav2Lip [36]	19.875	0.633	1.438/2.138	44.510	**8.774**	13.69	22.323	0.727	1.759/2.002	22.397	**9.032**
Audio2Head [51]	18.764	0.586	2.053/2.293	27.236	6.494	16.35	21.608	0.702	1.983/2.060	29.385	7.076
PC-AVS [65]	16.120	0.458	2.649/4.350	38.679	7.337	12.12	22.995	0.705	2.019/1.785	26.042	8.482
AVCT [52]	17.848	0.556	2.870/3.160	37.248	4.895	13.13	20.484	0.663	2.360/2.679	19.066	5.661
SadTalker [61]	19.042	0.606	2.038/2.335	39.308	7.065	14.25	21.701	0.702	1.995/2.147	14.261	7.414
IP-LAP [63]	19.832	0.627	2.140/2.116	46.502	4.156	17.34	22.615	0.731	1.951/1.938	19.281	3.456
TalkLip [48]	19.492	0.623	1.951/2.204	41.066	5.724	14.00	22.241	0.730	1.976/1.937	23.850	1.076
EAMM [20]	18.867	0.610	2.543/2.413	31.268	1.762	31.08	19.866	0.626	2.910/2.937	41.200	4.445
StyleTalk [31]	21.601	0.714	1.800/1.422	24.774	3.553	63.49	21.319	0.692	2.324/2.330	17.053	2.629
PD-FGC [47]	21.520	0.686	1.571/1.318	30.240	6.239	44.86	23.142	0.710	**1.626**/1.497	25.340	7.171
EAT [14]	20.007	0.652	1.750/1.668	21.465	7.984	64.40	22.076	0.719	2.176/1.781	28.759	7.493
EDTalk-A	**21.628**	**0.722**	**1.537**/**1.290**	**17.698**	8.115	**67.32**	**25.156**	**0.811**	1.676/**1.315**	**13.785**	7.642
EDTalk-V	**22.771**	**0.769**	**1.102**/**1.060**	**15.548**	6.889	**68.85**	**26.504**	**0.845**	**1.197**/**1.111**	**13.172**	6.732
GT	1.000	1.000	0.000/0.000	0.000	7.364	79.65	1.000	1.000	0.000/0.000	0.000	7.721

Table 1: Quantitative comparisons with state-of-the-art methods.

Comparison Setting. We compare our method with: (a) emotion-agnostic talking face generation methods: MakeItTalk [66], Wav2Lip [36], Audio2Head [51], PC-AVS [65], AVCT [52], SadTalker [61], IP-LAP [63], TalkLip [48]. (b) Emotional talking face generation methods: EAMM [20], StyleTalk [31], PD-FGC [47], EMMN [45], EAT [14], EmoGen [15]. Different from previous work, EDTalk encapsulates the entire face generation process without any other sources (e.g. poses [14, 20], 3DMM [31, 57], phoneme [31, 52]) and pre-processing operations during inference, which facilitates the application. We evaluate our model in both audio-driven setting (EDTalk-A) and video-driven setting (EDTalk-V) *w.r.t.* (i) generated video quality using PSNR, SSIM [55] and FID [38]. (ii) audio-visual synchronization using Landmarks Distances on the Mouth (M-LMD) [6] and the confidence score of SyncNet [10]. (ii) emotional accuracy using Acc_{emo} calculated by pretrained Emotion-Fan [32] and Landmarks Distances on the Face (F-LMD). Partial results are moved to supplementary material due to limited space.

4.2 Experimental Results

Quantitative Results. The quantitative results are presented in Tab. 1, where our EDTalk-A and EDTalk-V achieve the best performance across most metrics, except $\text{Sync}_{\text{conf}}$. Wav2Lip pretrains their SyncNet discriminator on a large dataset [1], which might lead the model to prioritize achieving a higher $\text{Sync}_{\text{conf}}$ over optimizing visual performance. It is evident in the blur mouths generated by Wav2Lip and inferior M-LMD score to our method.

Qualitative Results. Fig. 4 demonstrates comparison of visual results. TalkLip and IP-LAP struggle to generate accurate lip motions. Despite elevated lip-synchronization of SadTalker, they can only produce slight lip motions with closed mouth and are also bothered by jitter between frames. StyleHEAT generates accurate mouth shape driven by Mouth GT video instead of audio but suffers from incorrect head pose and identity loss. This issue also plagues EmoGen, EAMM and PD-FGC. Besides, EmoGen and EAMM fail to perform the desired

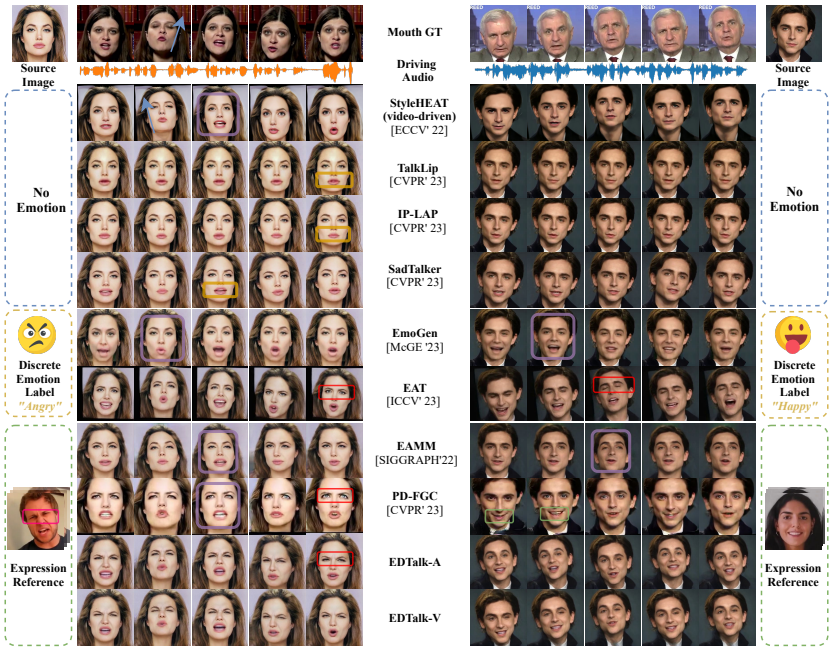

Fig. 4: Qualitative comparisons with state-of-the-art methods. See full comparison in supplementary material.

expression. Due to discrete emotion input, EAT cannot synthesize fine-grained expression like the narrowed eyes performed by expression reference. In the case of "happy", unexpected closed eyes and weird teeth are observed in EAT and PD-FGC, respectively. In contrast, both EDTalk-A and EDTalk-V excel in producing realistic expressions, precise lip synchronization and correct head poses.

Efficiency analysis. Our approach is highly efficient in terms of training time, required data and computational resources in decoupling spaces. In the mouth-pose decoupling stage, we solely utilize the HDTF dataset, containing **15.8 hours** of videos, for the decoupling. Training with a batch size of 4 on **two 3090 GPUs** for **4k** iterations achieves state-of-the-art performance, which takes about one hour. In contrast, DPE is trained on the VoxCeleb dataset, which comprises **351 hours** of video, for **100K** iterations initially, then an additional **50K** iterations with a batch size of 32 on

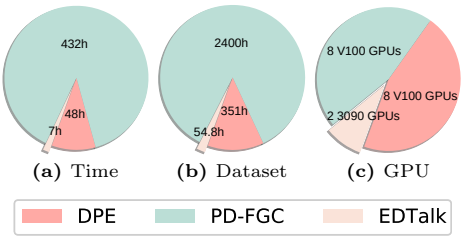

Fig. 5: Resources for training.

8 V100 GPUs, which takes over 2 days. Besides, they need to train two task-specific generators for expression and pose. Similarly, PD-FGC takes **2 days** on **4 Tesla V100 GPUs** for lip, and another **2 days on 4 Tesla V100 GPUs** for pose decoupling. It significantly exceeds our computational resources and training time. In the expression decouple stage, we train our model on MEAD and HDTF dataset (total **54.8 hours** of videos) for 6 hours. On the other hand, PD-FGC decouples expression space on Voxceleb2 dataset (**2400 hours**) by discorelation loss for 2 weeks. The visualization in Fig. 5 allows for a more intuitive comparison of the differences between the different methods concerning required training time, training data, and computational arithmetic.

4.3 Ablation Study

Latent space. To analyze the contributions of our key designs on obtaining the disentangled latent spaces, we conduct an ablation study with two variants: (1) remove base banks (**w/o Bank**). (2) remove orthogonal constraint (**w/o Orthogonal**). Fig. 6 presents our ablation study results on video-driven and audio-driven settings, respectively. Since **w/o Bank** struggles to decouple different latent spaces, *only exp* fails to extract the emotional expression. Additionally, without the visual information stored in banks, the quality of the generated full frame is poor. Although **w/o Orthogonal** improves the image quality through vision-rich banks, due to the lack of orthogonality constraints on the base, it interferes with different spaces, resulting in less obvious generated emotions. The Full Model achieves the best performance in both aspects.

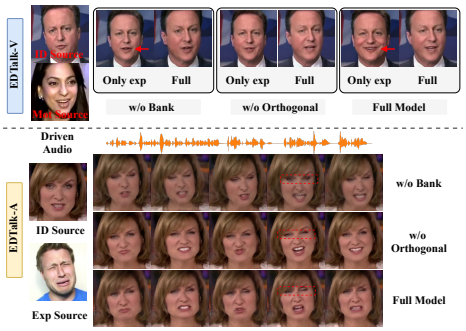

Fig. 6: Ablation results.

5 Conclusion

This paper introduces EDTalk, a novel system designed to efficiently disentangle facial components into latent spaces, enabling fine-grained control for talking head synthesis. The core insight is to represent each space with orthogonal bases stored in dedicated banks. We propose an efficient training strategy that autonomously allocates spatial information to each space, eliminating the necessity for external or prior structures. By integrating these spaces, we enable audio-driven talking head generation through a lightweight Audio-to-Motion module. Experiments showcase the superiority of our method in achieving disentangled and precise control over diverse facial motions. We provide more discussion about the limitations and ethical considerations in the supplementary material.

Acknowledgements

This work was supported by National Natural Science Foundation of China (NSFC, NO. 62102255), NetEase Fuxi Lab Industry-University Collaboration Research Funding.

References

1. Afouras, T., Chung, J.S., Senior, A., Vinyals, O., Zisserman, A.: Deep audio-visual speech recognition. IEEE transactions on pattern analysis and machine intelligence **44**(12), 8717–8727 (2018)
2. Blanz, V., Vetter, T.: A morphable model for the synthesis of 3d faces. In: Proceedings of the 26th annual conference on Computer graphics and interactive techniques. pp. 187–194 (1999)
3. Bregler, C., Covell, M., Slaney, M.: Video rewrite: Driving visual speech with audio. In: Seminal Graphics Papers: Pushing the Boundaries, Volume 2, pp. 715–722 (2023)
4. Chen, L., Cui, G., Liu, C., Li, Z., Kou, Z., Xu, Y., Xu, C.: Talking-head generation with rhythmic head motion. In: European Conference on Computer Vision. pp. 35–51. Springer (2020)
5. Chen, L., Li, Z., Maddox, R.K., Duan, Z., Xu, C.: Lip movements generation at a glance. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 520–535 (2018)
6. Chen, L., Maddox, R.K., Duan, Z., Xu, C.: Hierarchical cross-modal talking face generation with dynamic pixel-wise loss. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 7832–7841 (2019)
7. Chen, L., Wu, Z., Li, R., Bao, W., Ling, J., Tan, X., Zhao, S.: Vast: Vivify your talking avatar via zero-shot expressive facial style transfer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2977–2987 (2023)
8. Chung, J.S., Nagrani, A., Zisserman, A.: Voxceleb2: Deep speaker recognition. arXiv preprint arXiv:1806.05622 (2018)
9. Chung, J.S., Zisserman, A.: Lip reading in the wild. In: Computer Vision–ACCV 2016: 13th Asian Conference on Computer Vision, Taipei, Taiwan, November 20-24, 2016, Revised Selected Papers, Part II 13. pp. 87–103. Springer (2017)
10. Chung, J.S., Zisserman, A.: Out of time: automated lip sync in the wild. In: Computer Vision–ACCV 2016 Workshops: ACCV 2016 International Workshops, Taipei, Taiwan, November 20-24, 2016, Revised Selected Papers, Part II 13. pp. 251–263. Springer (2017)
11. Daněček, R., Black, M.J., Bolkart, T.: Emoca: Emotion driven monocular face capture and animation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20311–20322 (2022)
12. Das, D., Biswas, S., Sinha, S., Bhowmick, B.: Speech-driven facial animation using cascaded gans for learning of motion and texture. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16. pp. 408–424. Springer (2020)
13. Ekman, P., Friesen, W.V.: Facial action coding system. Environmental Psychology & Nonverbal Behavior (1978)
14. Gan, Y., Yang, Z., Yue, X., Sun, L., Yang, Y.: Efficient emotional adaptation for audio-driven talking-head generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 22634–22645 (2023)

15. Goyal, S., Bhagat, S., Uppal, S., Jangra, H., Yu, Y., Yin, Y., Shah, R.R.: Emotionally enhanced talking face generation. In: Proceedings of the 1st International Workshop on Multimedia Content Generation and Evaluation: New Methods and Practice. pp. 81–90 (2023)
16. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9729–9738 (2020)
17. Hong, F.T., Zhang, L., Shen, L., Xu, D.: Depth-aware generative adversarial network for talking head video generation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3397–3406 (2022)
18. Hsu, W.N., Bolte, B., Tsai, Y.H.H., Lakhotia, K., Salakhutdinov, R., Mohamed, A.: Hubert: Self-supervised speech representation learning by masked prediction of hidden units. IEEE/ACM Transactions on Audio, Speech, and Language Processing **29**, 3451–3460 (2021)
19. Huang, X., Belongie, S.: Arbitrary style transfer in real-time with adaptive instance normalization. In: Proceedings of the IEEE international conference on computer vision. pp. 1501–1510 (2017)
20. Ji, X., Zhou, H., Wang, K., Wu, Q., Wu, W., Xu, F., Cao, X.: Eamm: One-shot emotional talking face via audio-based emotion-aware motion model. In: ACM SIGGRAPH 2022 Conference Proceedings. pp. 1–10 (2022)
21. Ji, X., Zhou, H., Wang, K., Wu, W., Loy, C.C., Cao, X., Xu, F.: Audio-driven emotional video portraits. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 14080–14089 (2021)
22. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14. pp. 694–711. Springer (2016)
23. Khakhulin, T., Sklyarova, V., Lempitsky, V., Zakharov, E.: Realistic one-shot mesh-based head avatars. In: European Conference on Computer Vision. pp. 345–362. Springer (2022)
24. Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., Krishnan, D.: Supervised contrastive learning. Advances in neural information processing systems **33**, 18661–18673 (2020)
25. Kim, T., Vossen, P.: Emoberta: Speaker-aware emotion recognition in conversation with roberta. arXiv preprint arXiv:2108.12009 (2021)
26. Li, D., Zhao, K., Wang, W., Peng, B., Zhang, Y., Dong, J., Tan, T.: Ae-nerf: Audio enhanced neural radiance field for few shot talking head synthesis. arXiv preprint arXiv:2312.10921 (2023)
27. Liang, B., Pan, Y., Guo, Z., Zhou, H., Hong, Z., Han, X., Han, J., Liu, J., Ding, E., Wang, J.: Expressive talking head generation with granular audio-visual control. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3387–3396 (2022)
28. Liu, T., Chen, F., Fan, S., Du, C., Chen, Q., Chen, X., Yu, K.: Anitalker: Animate vivid and diverse talking faces through identity-decoupled facial motion encoding. arXiv preprint arXiv:2405.03121 (2024)
29. Liu, X., Xu, Y., Wu, Q., Zhou, H., Wu, W., Zhou, B.: Semantic-aware implicit neural audio-driven video portrait generation. In: European Conference on Computer Vision. pp. 106–125. Springer (2022)
30. Ma, Y., Wang, S., Ding, Y., Ma, B., Lv, T., Fan, C., Hu, Z., Deng, Z., Yu, X.: Talkclip: Talking head generation with text-guided expressive speaking styles. arXiv preprint arXiv:2304.00334 (2023)

31. Ma, Y., Wang, S., Hu, Z., Fan, C., Lv, T., Ding, Y., Deng, Z., Yu, X.: Styletalk: One-shot talking head generation with controllable speaking styles. arXiv preprint arXiv:2301.01081 (2023)
32. Meng, D., Peng, X., Wang, K., Qiao, Y.: Frame attention networks for facial expression recognition in videos. In: 2019 IEEE international conference on image processing (ICIP). pp. 3866–3870. IEEE (2019)
33. Pang, Y., Zhang, Y., Quan, W., Fan, Y., Cun, X., Shan, Y., Yan, D.m.: Dpe: Disentanglement of pose and expression for general video portrait editing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 427–436 (2023)
34. Park, S.J., Kim, M., Hong, J., Choi, J., Ro, Y.M.: Synctalkface: Talking face generation with precise lip-syncing via audio-lip memory. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 36, pp. 2062–2070 (2022)
35. Pataranutaporn, P., Danry, V., Leong, J., Punpongsanon, P., Novy, D., Maes, P., Sra, M.: Ai-generated characters for supporting personalized learning and well-being. Nature Machine Intelligence **3**(12), 1013–1022 (2021)
36. Prajwal, K., Mukhopadhyay, R., Namboodiri, V.P., Jawahar, C.: A lip sync expert is all you need for speech to lip generation in the wild. In: Proceedings of the 28th ACM International Conference on Multimedia. pp. 484–492 (2020)
37. Rezende, D., Mohamed, S.: Variational inference with normalizing flows. In: International conference on machine learning. pp. 1530–1538. PMLR (2015)
38. Seitzer, M.: pytorch-fid: FID Score for PyTorch. `https://github.com/mseitzer/pytorch-fid` (August 2020), version 0.3.0
39. Shen, S., Li, W., Zhu, Z., Duan, Y., Zhou, J., Lu, J.: Learning dynamic facial radiance fields for few-shot talking head synthesis. In: European Conference on Computer Vision. pp. 666–682. Springer (2022)
40. Shen, S., Zhao, W., Meng, Z., Li, W., Zhu, Z., Zhou, J., Lu, J.: Difftalk: Crafting diffusion models for generalized audio-driven portraits animation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1982–1991 (2023)
41. Siarohin, A., Lathuilière, S., Tulyakov, S., Ricci, E., Sebe, N.: First order motion model for image animation. Advances in neural information processing systems **32** (2019)
42. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
43. Sinha, S., Biswas, S., Yadav, R., Bhowmick, B.: Emotion-controllable generalized talking face generation. In: International Joint Conference on Artificial Intelligence. IJCAI (2021)
44. Song, Y., Zhu, J., Li, D., Wang, A., Qi, H.: Talking face generation by conditional recurrent adversarial network. In: Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence (Aug 2019). `https://doi.org/10.24963/ijcai.2019/129`, `http://dx.doi.org/10.24963/ijcai.2019/129`
45. Tan, S., Ji, B., Pan, Y.: Emmn: Emotional motion memory network for audio-driven emotional talking face generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 22146–22156 (2023)
46. Thies, J., Elgharib, M., Tewari, A., Theobalt, C., Nießner, M.: Neural voice puppetry: Audio-driven facial reenactment. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVI 16. pp. 716–731. Springer (2020)

47. Wang, D., Deng, Y., Yin, Z., Shum, H.Y., Wang, B.: Progressive disentangled representation learning for fine-grained controllable talking head synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 17979–17989 (2023)
48. Wang, J., Qian, X., Zhang, M., Tan, R.T., Li, H.: Seeing what you said: Talking face generation guided by a lip reading expert. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14653–14662 (2023)
49. Wang, J., Zhao, K., Zhang, S., Zhang, Y., Shen, Y., Zhao, D., Zhou, J.: 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. pp. 13844–13853 (2023)
50. Wang, K., Wu, Q., Song, L., Yang, Z., Wu, W., Qian, C., He, R., Qiao, Y., Loy, C.C.: Mead: A large-scale audio-visual dataset for emotional talking-face generation. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXI. pp. 700–717. Springer (2020)
51. Wang, S., Li, L., Ding, Y., Fan, C., Yu, X.: Audio2head: Audio-driven one-shot talking-head generation with natural head motion. In: International Joint Conference on Artificial Intelligence. IJCAI (2021)
52. Wang, S., Li, L., Ding, Y., Yu, X.: One-shot talking face generation from single-speaker audio-visual correlation learning. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 36, pp. 2531–2539 (2022)
53. Wang, Y., Yang, D., Bremond, F., Dantcheva, A.: Latent image animator: Learning to animate images via latent space navigation. In: International Conference on Learning Representations (2021)
54. Wang, Y., Boumadane, A., Heba, A.: A fine-tuned wav2vec 2.0/hubert benchmark for speech emotion recognition, speaker verification and spoken language understanding. arXiv preprint arXiv:2111.02735 (2021)
55. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE Transactions on Image Processing (2004)
56. Yang, K., Chen, K., Guo, D., Zhang, S.H., Guo, Y.C., Zhang, W.: Face2face ρ: Real-time high-resolution one-shot face reenactment. In: European conference on computer vision. pp. 55–71. Springer (2022)
57. Yin, F., Zhang, Y., Cun, X., Cao, M., Fan, Y., Wang, X., Bai, Q., Wu, B., Wang, J., Yang, Y.: Styleheat: One-shot high-resolution editable talking face generation via pre-trained stylegan. In: European conference on computer vision. pp. 85–101. Springer (2022)
58. Yu, Z., Yin, Z., Zhou, D., Wang, D., Wong, F., Wang, B.: Talking head generation with probabilistic audio-to-visual diffusion priors. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7645–7655 (2023)
59. Zakharov, E., Shysheya, A., Burkov, E., Lempitsky, V.: Few-shot adversarial learning of realistic neural talking head models. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9459–9468 (2019)
60. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 586–595 (2018)
61. Zhang, W., Cun, X., Wang, X., Zhang, Y., Shen, X., Guo, Y., Shan, Y., Wang, F.: Sadtalker: Learning realistic 3d motion coefficients for stylized audio-driven single image talking face animation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8652–8661 (2023)

62. Zhang, Z., Li, L., Ding, Y., Fan, C.: Flow-guided one-shot talking face generation with a high-resolution audio-visual dataset. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3661–3670 (2021)
63. Zhong, W., Fang, C., Cai, Y., Wei, P., Zhao, G., Lin, L., Li, G.: Identity-preserving talking face generation with landmark and appearance priors. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9729–9738 (2023)
64. Zhou, H., Liu, Y., Liu, Z., Luo, P., Wang, X.: Talking face generation by adversarially disentangled audio-visual representation. In: Proceedings of the AAAI conference on artificial intelligence. vol. 33, pp. 9299–9306 (2019)
65. Zhou, H., Sun, Y., Wu, W., Loy, C.C., Wang, X., Liu, Z.: Pose-controllable talking face generation by implicitly modularized audio-visual representation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4176–4186 (2021)
66. Zhou, Y., Han, X., Shechtman, E., Echevarria, J., Kalogerakis, E., Li, D.: Makelttalk: speaker-aware talking-head animation. ACM Transactions On Graphics (TOG) **39**(6), 1–15 (2020)