FlowVQTalker: High-Quality Emotional Talking Face Generation through Normalizing Flow and Quantization

Shuai Tan, Bin Ji, and Ye Pan[*]
Shanghai Jiao Tong University
{tanshuai0219,bin.ji,whitneypanye}@sjtu.edu.cn

Abstract

*Generating emotional talking faces is a practical yet challenging endeavor. To create a lifelike avatar, we draw upon two critical insights from a human perspective: 1) The connection between audio and the non-deterministic facial dynamics, encompassing expressions, blinks, poses, should exhibit synchronous and one-to-many mapping. 2) Vibrant expressions are often accompanied by emotion-aware high-definition (HD) textures and finely detailed teeth. However, both aspects are frequently overlooked by existing methods. To this end, this paper proposes using normalizing **Flow** and **V**ector-**Q**uantization modeling to produce emotional **talk**ing faces that satisfy both insights concurrently (**FlowVQTalker**). Specifically, we develop a flow-based coefficient generator that encodes the dynamics of facial emotion into a multi-emotion-class latent space represented as a mixture distribution. The generation process commences with random sampling from the modeled distribution, guided by the accompanying audio, enabling both lip-synchronization and the uncertain nonverbal facial cues generation. Furthermore, our designed vector-quantization image generator treats the creation of expressive facial images as a code query task, utilizing a learned codebook to provide rich, high-quality textures that enhance the emotional perception of the results. Extensive experiments are conducted to showcase the effectiveness of our approach.*

1. Introduction

Talking face generation has garnered growing interest due to its immense potential in various contexts, including virtual reality, filmmaking, and online education [34]. While existing research has made significant strides in improving lip-synchronization [4, 50, 51, 54], a notable oversight is the neglect of expressive facial expressions and diverse head poses, which are integral components of creating a lifelike and captivating avatar [46]. Consequently, one can readily distinguish such avatars from real humans.

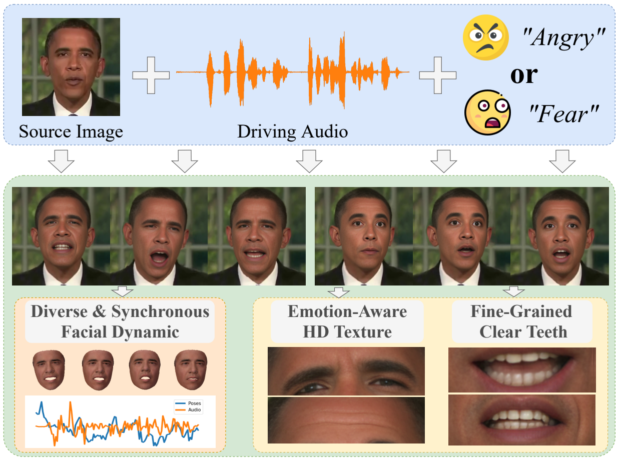


Figure 1. Example animations produced by FlowVQTalker. Given a source image and a driving audio, FlowVQTalker creates talking face video, complete with ***diverse and synchronous facial dynamic***, ***emotion-aware HD texture*** and ***fine-grained clear teeth***.

To address this, we initially identify two vital observations for natural-looking talking heads from a human perspective: 1) In reality, non-verbal facial cues exhibit inherent variability, rendering them non-deterministic in nature [45]. Therefore, the mapping from input audio to generated video constitutes a one-to-many relationship [49], where one audio clip can manifest in multiple plausible visual results owing to the fluid emotions, blinks, and head poses. 2) Authentic expressions not only retain the identity of source image but also feature intricate textures, such as wrinkles that intensify the expressiveness of the desired emotion. As depicted in Fig. 1, vertical lines between the eyebrows commonly accompany anger, while horizontal wrinkles on the forehead are associated with fear or surprise [10]. Furthermore, a high-quality video should encompass clear teeth. The observations succinctly encapsulate the key considerations in the avatar generation process, guiding the direction of progress and enhancement towards

[*] Corresponding author.

more engaging and realistic talking face synthesis.

Recent efforts have focused on modeling facial expressions [47, 48] and head motions. [11, 35, 43] rely on discrete emotional labels to prompt emotions, while [16, 22, 25] introduce an emotion reference video to suggest the desired expression. [59, 60, 65] infer head poses from audio via a time series analysis model [23]. Despite their contributions in enhancing the expressiveness of generated avatars, these methods still exhibit certain limitations: 1) Since nonverbal facial dynamics exhibit weak correlations with audio, such deterministic models tend to produce fixed and unrealistic outputs, lacking the diversity. 2) Low-resolution image generator [16, 25, 46] struggle to capture emotion-aware textures and clear teeth. Additionally, they face challenges in maintaining consistency between the generated video and source image [31, 56]. As a result, these methods have not addressed the two core observations mentioned earlier.

In this paper, we propose a novel method called FlowVQTalker to generate vibrant emotional talking head videos that meet: (1) diverse outputs encompassing various facial dynamics that respond to the driving audio and the emotional context. (2) preservation of the identity information from the source image, accompanied by the presentation of rich emotion-aware textures and clear teeth to enhance expressive performance and video quality. FlowVQTalker is comprised of the Flow-based Coeff. Generator (FCG) and Vector-Quantized Image Generator (VQIG), which are interconnected via the coefficients of 3D Morphable Models (3DMM) [7]. Within the FCG, we devise ExpFlow and PoseFlow for expression and pose coefficient modeling, built upon the generative flow model [39]. To elaborate, ExpFlow establishes an invertible transformation between emotional expression coefficients and latent codes, which are then mapped into the Student's t mixture model (SMM). Each mixture component of the SMM encodes features for an emotion class, wherein latent codes within the same component represent the same emotion while differing in nonverbal facial cues. During inference, we stochastically sample latent codes from the corresponding SMM component, ensuring the ***diversity*** of generated expressions. Furthermore, the one-to-one and bijective relationship between expression coefficients and latent codes enables us to achieve ***emotion transfer*** by providing an emotion reference. PoseFlow employs a similar technique to ExpFlow but incorporates specific modifications for handling pose-related aspects, as detailed in Sec. 3.2.

On the other hand, VQIG approaches the synthesis of fine-grained textures and teeth from a fresh perspective. We regard image rendering as a code query task within a learned codebook. This perspective is inspired by the capabilities of the codebook in VQ-GAN [9], which excels at preserving emotion-aware texture information and supplying a wealth of visual elements for generating top-quality faces. However, the vanilla VQ-GAN framework grapples with preserving identity information when appearing frequent spatial transformations. To this end, we extract features from source image to complement the motion synthesis, facilitating ***high-fidelity*** and ***expressive*** talking faces creation.

In summary, our contributions are outlined as follows:

- We introduce FlowVQTalker, a system capable of generating emotional talking face videos with diverse facial dynamics and fine-grained expressions.
- We harness the generative flow model to forecast nondeterministic and realistic coefficients. To the best of our knowledge, we are the pioneers in applying normalizing flow for emotional talking face generation.
- The visual codebook within our proposed VQIG enriches the textures with emotion-aware HD details, thereby enhancing expressiveness and elevating video quality.
- Extensive experiments demonstrate that our FlowVQTalker outperforms the competing methods in both quantitative and qualitative evaluation.

2. Related Work

2.1. Audio-Driven Talking Face Generation

Existing audio-driven talking face generation methods can be broadly categorized into reconstruction-based methods and intermediate representation-based methods. On the one hand, the former [4, 44, 71] typically involve mapping inputs from different modalities (e.g., audio and images) to corresponding features using encoders. Subsequently, these features are decoded to produce the talking faces. For example, Prajwal *et al.* [37] train their Wav2Lip following adversarial process [27], where the generator, based on an encoder-decoder architecture, reconstructs videos from the extracted features, and the lip-sync discriminator [6] is employed to improve lip-synchronization. On the other hand, the intermediate representations, like landmarks [5, 66, 70, 75] and dense motion fields [59, 60], are leveraged to bridge the modality gap. Zhou *et al.* [74] proposes a two-stage model to predict facial landmarks from audio, which subsequently serve as a condition for video synthesis. In contrast, we adopt 3D Morphable Model (3DMM) [7] as the bridge, as it offers better decoupling and control over expression, pose, and identity. Moreover, these methods neglect incorporation of expressive emotions into the generated faces, which is the main focus of our work.

2.2. Emotional Talking Face Generation

In the pursuit of achieving lifelike and emotionally expressive talking face generation, an increasing number of studies are incorporating emotion as a crucial element. One prevalent category of emotion sources in current methods is one-hot labels [8, 11, 18, 32, 33, 36, 43, 55]. Wang *et al.* [57] release an emotional audio-visual dataset MEAD and aligns neutral images with emotional states through the guidance of one-hot emotion labels using a U-Net. However, such deterministic models are prone to producing fixed expressions. Instead, our approach involves mapping different expressions into the latent distribution of a mixture model through normalizing flow [15, 39]. This enables us to sample a range of diverse expressions from the modeled distributions during inference. Recent works [16, 22, 52, 64] focus on emulating the speaking styles of given reference, thereby conveying more diverse expressions. Ma *et al.* [25], for instance, extract style codes from the 3DMM expression coefficients and generate lip movements synchronized with audio. Leveraging the forward process of normalizing flow, our method readily identifies the precise latent code corresponding to the given coefficients in the modeled distribution, thus achieving controllable emotional transfer.

2.3. Vector-Quantized Codebook

Vector-Quantized Network [9, 30] learns a codebook to store quantized features extracted from an autoencoder, significantly enhancing image modeling. Due to its effectiveness in replacing extracted features with quantized ones, this approach has demonstrated its remarkable potential in image restoration [12, 62, 73] and talking face generation [53, 58, 63]. Ng *et al.* [29] store facial motion in a discrete codebook and generate potential responsive motions of listeners during conversations. Wang *et al.* [58] learn position-invariant quantized local patch representations and employ a transformer to achieve face reenactment. However, both methods construct speaker-specific codebooks, which face challenges to generalize on arbitrary identities and facial motions. In contrast, our approach utilizes a generic codebook capable of representing a wide range of identities and facial expressions, enabling the production of high-quality talking face videos with rich emotional facial textures. To the best of our knowledge, we are the pioneers in employing normalizing flow and vector-quantized codebooks for speech-driven emotional facial animation.

3. Method

3.1. Overview

Fig. 2a illustrates the pipeline of FlowVQTalker, which is designed to create emotional talking head videos based on a source image I_0, driving audio a and emotion label e. We also achieve emotion transfer by introducing an emotion reference I_e^r. Basically, we leverage the expression coefficients $\beta \in \mathbb{R}^{64}$ and pose coefficients $\rho \in \mathbb{R}^6$ of 3D Morphable Models (3DMM) [3] to represent facial dynamics. Specifically, we input $\{(\beta_0, \rho_0), a, e\}$ or $\{(\beta_0, \rho_0), a, \beta_e^r\}$ (for emotion transfer) into Flow-based Coeff. Generator (Sec. 3.2), comprising ExpFlow and PoseFlow, to predict emotional facial dynamics $(\hat{\beta}_e, \hat{\rho}_e)$. By integrating the designed Vector-Quantized Image Generator (Sec. 3.3), we generate emotional talking head frames $\hat{I}_e$ from the generated $(\hat{\beta}_e, \hat{\rho}_e)$ and the source image I_0. The following subsections will delve into the finer details of each module.

3.2. Flow-based Coeff. Generator

Preliminary of Generative Flow Model. Normalizing flow [39] serves as a generative model with a primary advantage: it can effectively model a complex distribution $\mathcal{X}$ by leveraging a simple, fixed base distribution $\mathcal{Z}$ through an invertible and differentiable nonlinear transformation f. This reversibility of f implies that $z \in \mathcal{Z}$ can be readily obtained from $x \in \mathcal{X}$ using an inverse process denoted as f^{-1}, constituting a *normalizing flow*. When dealing with highly complex distributions, it is often necessary to employ a series of multiple flow steps $\{f_n\}_{n=1}^K$ to achieve the desired distribution modeling: $f = f_K \circ \cdots \circ f_2 \circ f_1$. Formulaically, for a given complex distribution $\mathcal{X} \sim p_{\mathcal{X}}$, a simple distribution $\mathcal{Z} \sim p_{\mathcal{Z}}$ and hidden distributions $\mathcal{H}_n$, the transformations among these can be represented as:

$$z \xleftrightarrow{f_1} h_1 \xleftrightarrow{f_2} h_2 \cdots \xleftrightarrow{f_K} x \tag{1}$$

$$x = f(z) = f_K(f_{K-1}(...f_1(z))) \tag{2}$$

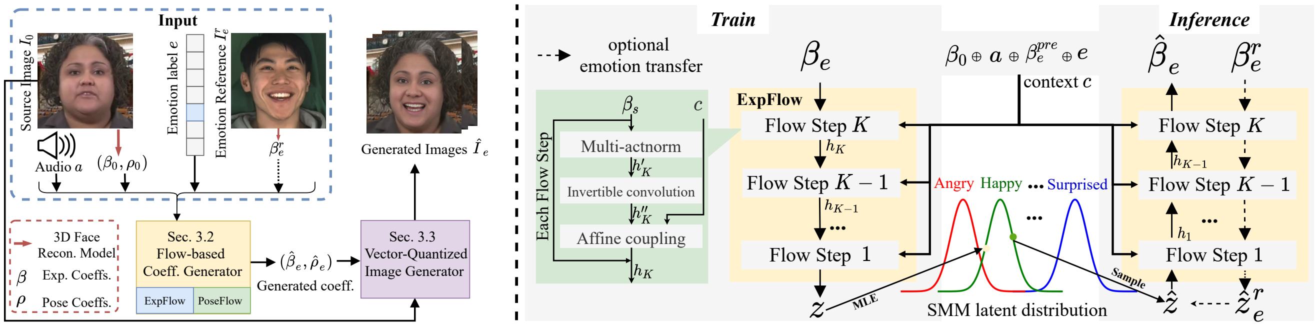
$$z = f^{-1}(x) = f_1^{-1}(f_2^{-1}(...f_K^{-1}(x))) \tag{3}$$

Given $\left(\frac{\partial z_k}{\partial z_{k-1}}\right)$ as the Jacobian matrix of f_n^{-1} at x, the log-likelihood [2] is formulated using maximum likelihood:

$$\log_{\mathcal{X}}(x) = \log_{\mathcal{Z}}(z) + \sum_{k=1}^{K} \log \left| \det \left(\frac{\partial z_k}{\partial z_{k-1}} \right) \right|, \tag{4}$$

ExpFlow. We apply normalizing flow [39] to produce the expression coefficients of talking face with diverse facial emotion dynamics. However, several non-trivial challenges have emerged: 1) Calculating $\det\left(\frac{\partial z_k}{\partial z_{k-1}}\right)$ comes with computational complexity that approaches $\mathcal{O}(D^3)$, which is intractable for large input dimension D. 2) Encoding various emotional coefficients into the latent space and further sampling the specified latent code z has been minimally explored. 3) The scarcity of emotional audio-visual dataset poses difficulties for modeling mixture distributions. To tackle these challenges, we introduce ExpFlow, which not only establishes a more efficient architecture but also leverages a mixture model designed for few-shot learning.

Fig. 2b illustrates the framework of ExpFlow built on [13, 15, 21]. The ground truth (GT) emotional coefficients $\beta_e^t \in \mathbb{R}^{64}$ (for simplicity, we omit time t in the following) and conditional context c pass through K flow steps to yield latent code $z \in \mathbb{R}^{64}$. In our work, the context



(a) The pipeline of our proposed FlowVQTalker

(b) The framework of ExpFlow and train/inference workflow.

Figure 2. The overview of our proposed FlowVQTalker. (a) Main pipeline. Given a source image I_0, audio a and emotion label e, Flow-based Coeff. Generator (Sec. 3.2) generates synchronized emotional coefficients $(\hat{\beta}_e, \hat{\rho}_e)$. Vector-Quantized Image Generator (Sec. 3.3) further produces emotional talking face frames $\hat{I}_e$. (b) The framework of ExpFlow and train/inference workflow. ExpFlow is composed of K flow steps, each containing three subsections: multi-actnorm, invertible convolution, and an affine coupling layer. Based on the property of normalizing flow [40], our proposed ExpFlow is bijective.

c contains the source coefficient β_0 for preserving identity, driving audio a for lip motion guidance, the previous τ coefficients $\beta_e^{pre} = \beta_e^{t-\tau:t-1}$ for video frame consistency and emotion label e for specifying desired expression.

Each flow step of transformation $f_i^{-1}(\cdot)$ consists of multi-actnorm, invertible convolution and the affine coupling layer. A detailed schematic is provided in the supplementary (*Suppl*). For the multi-actnorm, given the mean μ and standard deviation δ for each set of emotional data, we implement it as an affine transformation: $h' = \frac{\beta-\mu}{\delta}$. In our case, we initialize μ and δ with the same parameters for each emotion and update them during training, which helps mitigate overfitting [15]. Next, ExpFlow introduces an invertible 1×1 convolution layer: $h'' = \mathbf{W} \cdot h'$, designed to handle potential channel variations. Following this, we utilize a coupling layer based on a transformer $\mathcal{F}$ to generate h from h'' and c. More specifically, we split h'' into h''_{h1} and h''_{h2}, where h''_{h2} is affinely transformed by $\mathcal{F}$ based on h''_{h1}:

$$t, s = \mathcal{F}(h''_{h1}, c); \qquad h = [h''_{h1}, (h''_{h2} + t) \odot s], \tag{5}$$

where t and s denote the transformation parameters. Thanks to the preserved h''_{h1}, we can maintain tractability in the reverse direction. To sum up, we have the capability to map β_e into the latent code z and predict coefficients based on a sampled code $\hat{z} \in p\mathcal{Z}$ as follows:

$$z = f^{-1}(\beta_e, c); \qquad \hat{\beta}_e = f(\hat{z}, c) \tag{6}$$

Furthermore, we mitigate the computational complexity from $\mathcal{O}(D^3)$ by streamlining the computation of the Jacobian determinant using a transformer $\mathcal{F}$, analyzed in [21].

Up to this point, the need for a suitable distribution model $p_{\mathcal{Z}}$ is paramount. To this end, we turn to Student's t Mixture Model (SMM) which encompasses a 'fat tail' of multivariate t-distribution, particularly effective when working with our relatively small datasets [1]. Concretely, if the outliers within the coefficient distribution cannot be adequately explained by the latent distribution, they exert an unbounded influence on the maximum likelihood process. Therefore, to mitigate the impact of these outlier data points, the t-distribution becomes our preferred choice. Given emotion label $e \in 1, \cdots, C$, mean μ_i, Σ_i and the degrees of freedom $\nu (> 0)$, a multivariate t-distribution and marginal distribution of z known as SMM are expressed as:

$$p_{\mathcal{Z}}(z \mid e = i) = t_\nu \left(z \mid \mu_i, \Sigma_i\right), \tag{7}$$

$$p_{\mathcal{Z}}(z) = \sum_{i=1}^{C} \pi_i t_\nu \left(z \mid \mu_i, \Sigma_i\right), \tag{8}$$

where π_i is the mixture coefficient. Here we hypothesise that all categories of emotions have the same proportion, setting $\pi_i = \frac{1}{C}$. Following [15], the emotion-class-conditional likelihoods of β_e is:

$$p_{\beta_e \sim \mathcal{X}}(\beta_e \mid e = i) = t_\nu \left(z \mid \mu_i, \Sigma_i\right) \cdot \prod_{k=1}^{K} \left| \det \left(\frac{\partial z_k}{\partial z_{k-1}} \right) \right| \tag{9}$$

To ensure that the mean vectors of different emotions are distinct from each other in the latent space, we randomly initialize the mean vectors μ of each emotion i from the standard normal distribution $\mu_i \sim \mathcal{N}(0, \mathbf{I})$ and maintain their covariance matrices as identity according to [15]. In combination with SMM, we can train our ExpFlow by minimizing the negative log-likelihood loss (MLE):

$$\mathcal{L}_{\text{exp}} = -\sum_{t=0}^{T-1} \log p_{\mathcal{Z}} \left(f^{-1}(\beta_e^t, c) \mid e^t\right), \tag{10}$$

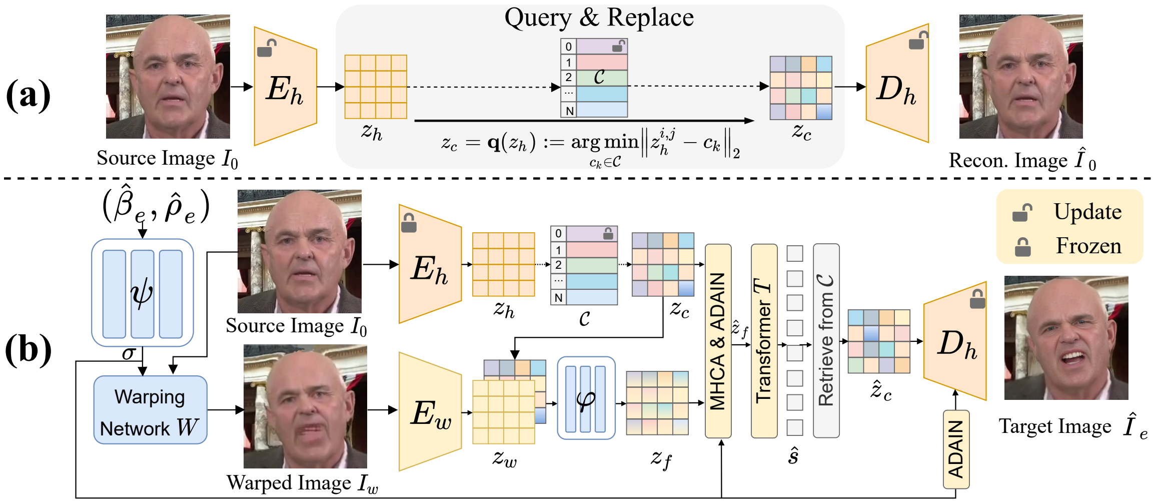


Figure 3. The structure of Vector-Quantized Image Generator (VQIG). (a) Codebook. Initially, we train a high-quality image encoder E_h, codebook $\mathcal{C}$ and image decoder D_h using a self-reconstruction strategy. Once converged, we freeze the parameters of E_h, $\mathcal{C}$ and D_h. (b) VQIG. I_0 are encoded as a discrete representation z_c to capture identity and texture information. $(\hat{\beta}_e, \hat{\rho}_e)$ generated by Sec. 3.2 are mapped into σ by ψ, which is then used to warp I_0 into I_w. An additional warped image encoder E_w is introduced to encode I_w as z_w. We employ φ to derive z_f from z_w and z_c. Along with z_c, both are fed into the multi-head cross-attention module (MHCA) and a transformer T. The generated vectors retrieve the best-matching features from the Codebook $\mathcal{C}$. The final result is rendered through D_h. It is worth noting that we introduce ADAIN at both feature and image levels to incorporate additional spatial information.

where T is the length of audio a. To guarantee video continuity, we introduce the consistency loss $\mathcal{L}_{\text{con}}$:

$$\mathcal{L}_{\text{con}} = \|(\beta_e^t - \beta_e^{t-1}) - (\hat{\beta}_e^t - \hat{\beta}_e^{t-1})\|_1, \quad (11)$$

where $\hat{\beta}_e$ is generated by the reverse process in Eq. (6).

During inference, we produce the coefficient sequence in an autoregressive fashion, with the current output $\hat{\beta}_e^t$ serving as the context for the subsequent iteration, as the previous frame β_e^{pre}. Furthermore, we have the option to replace the randomly sampled $\hat{z}$ with $\hat{z}_e^r = f^{-1}(\beta_e^r)$ to facilitate emotion transfer, where β_e^r are extracted from the emotion reference I_e^r, as depicted in 2a.

We notice that the dropout of β_e^{pre} in context c significantly influences emotion diversity (as confirmed in Fig. 6). Besides, while the 'fat tail' of the t-distribution has effectively addressed the outlier issue resulting from limited data, we have further improved performance through Manifold Projection [24]. Particularly, we apply the trained ExpFlow to obtain and store all $z \in \mathbb{R}^{64}$ from the training set as the prior $\mathcal{D} \in \mathbb{R}^{N \times 64}$. At inference time, for each randomly sampled $\hat{z}$, we find the $\mathcal{K}$ nearest points $\{\overline{z}_1, \cdots, \overline{z}_\mathcal{K}\}$ in $\mathcal{D}$. Subsequently, we substitute $\sum_{k=1}^{\mathcal{K}} w_k \cdot \overline{z}_k$ for $\hat{z}$, where the weights w_k are determined by minimizing $\min \|\hat{z} - \sum_{k=1}^{\mathcal{K}} w_k \cdot \overline{z}_k\|_2^2$, $\sum_{k=1}^{\mathcal{K}} w_k = 1$. In this way, we maintain the original diversity while enhancing robustness.

PoseFlow. PoseFlow is designed to generate a sequence of head poses based on audio and pose history. To achieve this, PoseFlow follows a similar framework as ExpFlow shown in Fig. 2b. However, there are several key modifications. First, given that the emotional dataset MEAD [57] used in our work, used in our work lacks pose information, we adapt the context c from $[\beta_0, a, \beta_e^{pre}, e]$ to $c_{\text{pose}} = [a, \rho^{pre}]$. The rationale for excluding the source head pose ρ_0 is that a single frame's pose does not convey identity information. Second, we employ a Gaussian distribution $\mathcal{N}(0, \mathbf{I})$ as our $p_\mathcal{Z}$ instead of SMM, as we disregard the influence of emotion on head pose. Third, given the pose flow step f_{pose}, we replace the loss function in Eq. (10) with $\mathcal{L}_{\text{pose}}$:

$$\mathcal{L}_{\text{pose}} = -\sum_{t=0}^{T-1} \log p_\mathcal{Z} \left(f_{pose}^{-1}(\rho^t, c_{pose})\right) \quad (12)$$

3.3. Vector-Quantized Image Generator

Upon obtaining the emotional coefficients $(\hat{\beta}_e, \hat{\rho}_e)$, we employ them to animate the source image I_0 with intricate textures, which are crucial for conveying desired emotions. Our key insight is to build a discrete codebook, which stores high-quality visual textures of face images including teeth, providing essential details to enhance the quality of the animated results. Leveraging this context-rich codebook, we introduce our Vector-Quantized Image Generator (VQIG) to enable high-fidelity and expressive image rendering.

Codebook. We draw inspiration from VQGAN [9], and train our texture-preserving codebook prior by self-reconstruction. Concretely, as illustrated in Fig. 3a, we initially apply a high-quality image encoder E_h to embed the source image $I_0 \in \mathbb{R}^{H\times W\times 3}$ into the latent vector $z_h \in \mathbb{R}^{m\times n\times d}$. Then we generate the vector-quantized representation z_c by involving an introduced codebook $\mathcal{C} = \{c_k \in \mathbb{R}^d\}_{k=1}^N$ and replacing z_c with the queried nearest code c_k in $\mathcal{C}$:

$$z_c = \mathbf{q}(z_h) := \underset{c_k\in\mathcal{C}}{\arg\min} \left\| z_h^{i,j} - c_k \right\|_2 . \tag{13}$$

Subsequently, an image decoder D_h is leveraged to reconstruct the input image $\hat{I}_0$. To jointly train the above modules in an end-to-end fashion, we adopt reconstruction loss $\mathcal{L}_{\text{rec}}$, perceptual loss $\mathcal{L}_{\text{per}}$ [17, 67] and adversarial loss $\mathcal{L}_{\text{adv}}$ [9]:

$$\mathcal{L}_{\text{rec}} = \|I_0 - \hat{I}_0\|_1; \qquad \mathcal{L}_{\text{per}} = \|\Phi(I_0) - \Phi(\hat{I}_0)\|_2^2; \tag{14}$$

$$\mathcal{L}_{\text{adv}} = \log D(I_0) + \log(1 - D(\hat{I}_0)), \tag{15}$$

where Φ denotes the feature extractor of VGG19 [42]. To update E_h and $\mathcal{C}$, we utilize code-level loss $\mathcal{L}_{\text{code}}$ and $\mathcal{L}_{\text{feat}}$:

$$\mathcal{L}_{\text{code}} = \|\text{sg}(z_h) - z_c\|_2^2; \qquad \mathcal{L}_{\text{feat}} = \|z_h - \text{sg}(z_c)\|_2^2, \tag{16}$$

where $\text{sg}(\cdot)$ donates the stop-gradient operator. Given the loss weights λs, the total loss $\mathcal{L}_{\text{tot}}$ is represented as:

$$\mathcal{L}_{\text{tot}} = \mathcal{L}_{\text{rec}} + \mathcal{L}_{\text{pre}} + \lambda_{\text{adv}}\mathcal{L}_{\text{adv}} + \mathcal{L}_{\text{code}} + \lambda_{\text{feat}}\mathcal{L}_{\text{feat}}. \tag{17}$$

VQIG. Once the codebook $\mathcal{C}$ is well-trained, we freeze the parameters of E_h, $\mathcal{C}$ and D_h, and devise Vector-Quantized Image Generator (VQIG) for image animation shown in Fig. 3b. Specifically, the source image I_0 is first transformed into $z_h = E_h(I_0)$ and quantize it as $z_c = \mathbf{q}(z_h)$, which delivers the identity and texture information. Considering that the predicted coefficients $(\hat{\beta}_e, \hat{\rho}_e)$ provide the motion guidance, we introduce a mapping network Φ [38] to generate motion descriptors $\sigma = \Phi(\hat{\beta}_e, \hat{\rho}_e)$, which are used to further warp I_0 as I_w via a warping network W. Subsequently, we extract z_w using warped image encoder E_w, which is fine-tuned from E_h during training VQIG. While I_w roughly achieves spatial transformation, it may struggle to preserve identity and texture information with detrimental artifacts. To this end, we compensate with z_c by combining it with z_w as z_f using a fuse network φ. To enhance the performance of emotion-aware texture, we employ a multi-head cross-attention mechanism (MHCA) [62] and an adaptive instance normalization (AdaIN) [14] operator to spatially fuse z_c and z_f, which helps to restore the face with fidelity and motion guidance, respectively. Subsequently, we insert a transformer T [73] to predict code sequence $\hat{s} \in \{0, \cdots, N-1\}^{m\cdot n}$, which retrieves the respective code items from learned codebook $\mathcal{C}$, forming quantized features $\hat{z}_c$. Through the decoder D_h, we generate high-fidelity and expressive face images $\hat{I}_e$ with clear teeth.

To train our VQIG, we adopt the two-stage train strategy [73]: code-level and image-level supervision. Firstly, we extract code sequence s and latent features z_c from GT frame and minimize the difference with the predicted ones:

$$\mathcal{L}_{\text{code}}^{\text{VQIG}} = \sum_{i=0}^{m\cdot n-1} -s_i \log(\hat{s}_i); \qquad \mathcal{L}_{\text{feat}}^{\text{VQIG}} = \|\hat{z}_f - z_c\|_2^2. \tag{18}$$

Besides, we incorporate the same image-level loss functions as Eq. (14) and Eq. (15).

4. Experiments

4.1. Experimental Settings

Datasets and Implementation Details. We train our framework on MEAD [57] and HDTF [69]. During training codebook, we additionally incorporate FFHQ dataset [19] to further improve face modeling capabilities and enhance the preservation of expressive textures. MEAD includes videos of 60 participants expressing 8 different emotions while speaking 30 sentences. HDTF collects various talking videos featuring over 300 identities from YouTube. We categorize the video clips in MEAD into eight emotion categories as originally specified and designate the video clips in HDTF as the ninth category, as they generally represent speaking styles closer to reality without pronounced emotions. FFHQ dataset comprises 70,000 high-quality face images. We crop and resize all data as a resolution of 512×512 and the latent vector dims are $m = n = 16, d = 256$. The codebook size and loss weights are set as $N = 1024$, $\lambda_{\text{adv}} = 0.8$, $\lambda_{\text{feat}} = 0.25$, respectively. Our method is implemented with PyTorch and trained using the Adam optimizer [20] on 4 NVIDIA GeForce GTX 3090.

Comparison Setting. We compare our method with: (a) emotion-agnostic talking face generation methods: Wav2Lip [37], PC-AVS [72], IP-LAP [70]; (b) emotional talking face generation methods: EAMM [16], EMMN [46], EAT [11], PD-FGC [52]. The former focuses on the synchronization between the generated lip motion and the input audio, which is validated by Landmarks Distances on the Mouth (M-LMD) [5] and the confidence score of SyncNet [6]. In contrast, the emotional talking head generation methods performs expressive expressions on the whole face, where we employ Facial Landmarks Distances (F-LMD) for evaluation. In addition, we fine-tune the Emotion-Fan [26] using MEAD dataset and measure the emotion accuracy (Acc_{emo}). Furthermore, SSIM [61] and FID [41] are adopted to assess image quality, while cumulative probability blur detection (CPBD) [28] is introduced to evaluate the clarity of the texture.

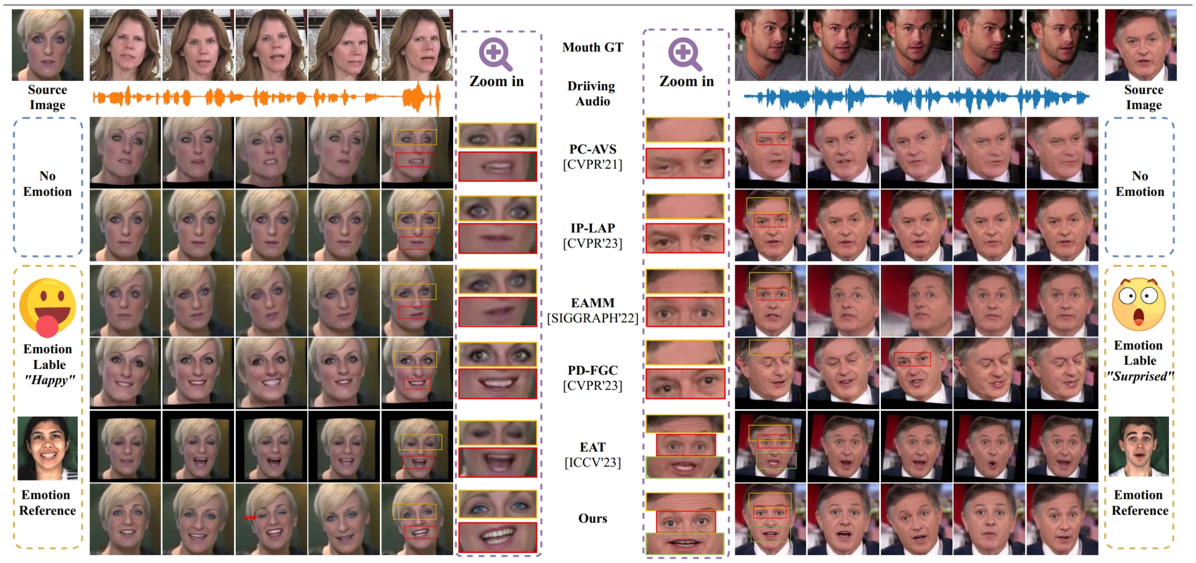


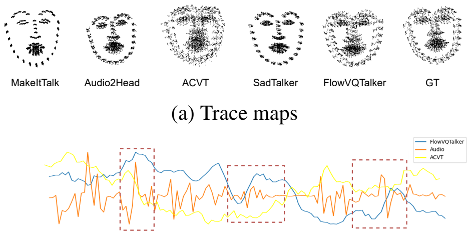
Figure 4. Qualitative comparisons with state-of-the-art methods. See full comparison in supplementary material (*Suppl*).

Method	**MEAD** [57]						**HDTF** [69]					**Emotion Input**		**Output**	
	SSIM↑	FID↓	M/ F-LMD↓	$\text{Sync}_{\text{conf}}$ ↑	CPBD↑	Acc_{emo} ↑	SSIM↑	FID↓	M/ F-LMD↓	$\text{Sync}_{\text{conf}}$ ↑	CPBD↑	Label	Reference	Diversity	HD
Wav2Lip [37]	0.648	28.924	2.294 / 2.234	**8.484**	0.104	17.64%	**0.742**	19.757	1.767 / 1.732	**9.073**	0.126	✗	✗	✗	✗
PC-AVS [72]	0.510	36.804	3.130 / 4.062	5.641	0.125	15.84%	0.690	17.617	1.637 / 2.217	8.520	0.119	✗	✗	✗	✗
IP-LAP [70]	0.641	26.823	2.303 / 2.205	3.371	0.116	17.61%	0.710	18.461	1.789 / **1.708**	3.357	0.142	✗	✗	✗	✗
EAMM [16]	0.621	26.478	2.624 / 2.762	1.594	0.106	43.68%	0.604	27.302	2.747 / 2.746	4.296	0.118	✗	✓	✗	✗
PD-FGC [52]	0.684	27.511	2.104 / 2.112	5.196	0.103	61.47%	0.692	16.929	1.720 / 1.966	7.321	0.128	✗	✓	✗	✗
EMMN [46]	0.675	22.895	2.439 / 2.851	5.125	0.116	58.53%	0.671	20.137	2.513 / 2.924	5.844	0.114	✓	✗	✗	✗
EAT [11]	0.684	19.836	2.056 / 2.284	6.533	0.120	65.83%	0.706	18.316	1.945 / 2.026	7.428	0.121	✓	✗	✗	✗
FlowVQTalker	**0.689**	**16.553**	**1.939 / 2.061**	5.901	**0.181**	**71.53%**	0.708	**15.165**	**1.643** / 1.958	6.766	**0.268**	✓	✓	✓	✓
GT	1.000	0.000	0.000 / 0.000	6.733	0.161	81.68%	1.000	0.000	0.000 / 0.000	7.728	0.238	-	-	-	-

Table 1. Quantitative comparisons with state-of-the-art methods. Supplementary material gives more quantitative comparison results.

4.2. Compare with other state-of-the-art methods

Talking Face Generation. Fig. 4 displays video frames generated by various SOTA methods (See *Suppl* for full comparison). PC-AVS and IP-LAP struggle with preserving identity information and lip synchronization, respectively. Additionally, both methods cannot generate videos with emotional expressions. Although EAMM resorts to a driving video as emotion guidance, it fails to perform vivid expression on the whole face. PD-FGC and EAT can predict happy faces, but PD-FGC's surprised faces lack clear expressions, and EAT faces issues with closed eyes. Our results demonstrate the preservation of identity information while accurately expressing corresponding emotions. Note that since $\hat{z}$ is randomly sampled from the modeled distribution, our method can randomly generate blinks (as pointed out by red arrow), contributing to expressive and dynamic talking faces. Furthermore, please see the zoom-in details, the compared methods struggle to synthesize fine-grained, emotion-aware textures and clear teeth. In contrast, our method excels at generating high-quality images, even when the source image is blurred. Tab. 1 shows that our FlowVQTalker achieves the best performance across most evaluation criteria. Wav2Lip [37] obtains the highest scores in $\text{Sync}_{\text{conf}}$ which even surpasses the ground truth (GT). We assume that Wav2Lip uses SyncNet confidence as a critical constraint using a SyncNet discriminator [6], which is reasonable to pursue a higher $\text{Sync}_{\text{conf}}$. We obtain similar scores to GT and lower M-LMD, demonstrating our ability to predict synchronized lip motions. Moreover, since Wav2Lip and IP-LAP only edit mouth region and keep other facial parts unchanged, they achieve highest SSIM and lowest F-LMD, respectively. Thanks to our texture-rich codebook, FlowVQTalker significantly outperforms SOTAs regarding CPBD on both datasets, which suggests that the details in our results are clearer and more visually appealing. Note that our emotion source can be emotion label for high-definition (HD) diverse facial emotion dynamics generation, or be an emotion reference for emotion transfer. In contrast, the compared methods can only accommodate one



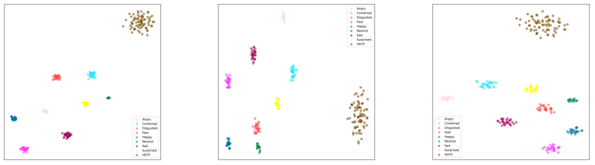
(a) Trace maps

(b) Correlation between audio and poses.

Figure 5. Comparison with SOTAs in terms of generated poses.

Metric/Method	PC-AVS	IP-LAP	EAMM	EAT	FlowVQTalker	GT
Lip-sync	3.89	3.91	3.43	3.94	**4.03**	4.88
Iamge-quality	3.24	3.83	3.46	3.72	**4.25**	4.67
Acc_{emo}	31.5%	54.9%	53.2%	52.4%	**60.6%**	76.4%

Table 2. User study results.



(a) GMM

(b) w/o data dropout

(c) Ours

Figure 6. Visualization of latent space.

of these inputs, resulting in a deterministic output.

Diversity. We present facial dynamics diversity and emotion transfer in supplementary video, encompassing expression, blink and pose. Here, we compare with MakeItTalk [75], Audio2Head [59], AVCT [60] and SadTalker [68] concerning pose diversity and correlation with audio, assessed using trace map [68] and correlation map [59]. Fig. 5a demonstrates that only AVCT performs comparable diversity with our FlowVQTalker, while ours achieves better synchronization than ACVT in Fig. 5b.

User Study. We conduct user study to evaluate our method from a human perspective. We recruit 20 participants (10 males/10 females) to score 120 videos (20 videos × (5 methods + GT)) from 1 (worst) to 5 (best) in terms of lip synchronization and image quality. They are also required to classify the emotion performed by videos. The results, detailed in Tab. 2, clearly illustrate the superiority of our method across all evaluated aspects with the aid of our human-like observations and corresponding solutions.

4.3. Ablation Study.

Ablation of ExpFlow. For ExpFlow, we delve into the impact of different settings on the latent space, a crucial element in emotional modeling and diversity. We examin two key variations: (1) **GMM**: replace SMM with Gaussian Mixture Model (GMM). (2) **w/o data dropout**: use SMM but exclude data dropout. Fig. 6 presents the visualization of the latent space. GMM tends to overfit and form clusters around specific data points within our relatively small dataset [57]. This leads to poor performance as the model is prone to sampling outliers. While SMM, with its long-tailed distributions, mitigates this issue, data dropout is also critical for constructing more resilient distributions. In addition, data dropout also improves the consistency between the generated motion and the context, achieving better synchronization of audio and lip motion as shown in *Suppl*.

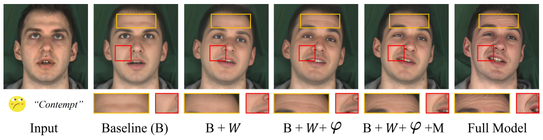


Figure 7. Visualization results of ablation study.

Ablation of VQIG. We conduct an ablation study on VQIG with following variants: We started with a **baseline (B)**, which retains only E_w, ADAIN and D_h. Subsequently, we add W, φ, MHCA and $\mathcal{C}$ in turn to verify the validity of each module, namely **B+W**, **B+W+φ**, **B+W+φ+M** and **B+W+φ+M+$\mathcal{C}$** (**Full Model**). The results presented in Fig. 7, where W initially warps the source image but lacks detail information, φ and MHCA effectively fuse identity, texture and spatial information, and the inclusion of $\mathcal{C}$ further improves the fidelity and clarity of the image.

5. Conclusion

In this paper, we introduce FlowVQTalker, a system capable of generating talking face with high-definition expression and non-deterministic facial dynamics, addressing both insights we set out. Within FlowVQTalker, flow-based coeff. generator establishes an invertible mapping between coefficients of 3DMM and a distribution model, allowing for random sampling to ensure diverse nonverbal facial expressions. Vector-quantized image generator resorts to a texture-rich codebook and synthesizes realistic videos with fine-grained details, such as emotion-aware wrinkles and clear teeth. We conduct comprehensive experiments to illustrate the superiority of our FlowVQTalker.

6. Acknowledgments

This work was supported by National Natural Science Foundation of China (NSFC, NO. 62102255), Shanghai Municipal Science and Technology Major Project (No. 2021SHZDZX0102). Tencent Industry-University Collaboration Research Funding (TencentJR2024IEG111). We would like to thank Xinya Ji, Yifeng Ma and Yuan Gan for their generous help.

References

[1] Simon Alexanderson and Gustav Eje Henter. Robust model training and generalisation with studentising flows. *arXiv preprint arXiv:2006.06599*, 2020. 4

[2] Anthony J Bell and Terrence J Sejnowski. An information-maximization approach to blind separation and blind deconvolution. *Neural computation*, 7(6):1129–1159, 1995. 3

[3] 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

[4] Lele Chen, Zhiheng Li, Ross K Maddox, Zhiyao Duan, and Chenliang Xu. Lip movements generation at a glance. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 520–535, 2018. 1, 2

[5] Lele Chen, Ross K Maddox, Zhiyao Duan, and Chenliang Xu. Hierarchical cross-modal talking face generation with dynamic pixel-wise loss. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7832–7841, 2019. 2, 6

[6] Joon Son Chung and Andrew Zisserman. 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*, pages 251–263. Springer, 2017. 2, 6, 7

[7] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 0–0, 2019. 2

[8] Sefik Emre Eskimez, You Zhang, and Zhiyao Duan. Speech driven talking face generation from a single image and an emotion condition. *arXiv: Audio and Speech Processing*, 2020. 2

[9] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2, 3, 5, 6

[10] Gary Faigin. *The artist's complete guide to facial expression*. Watson-Guptill, 2012. 1

[11] Yuan Gan, Zongxin Yang, Xihang Yue, Lingyun Sun, and Yi Yang. Efficient emotional adaptation for audio-driven talking-head generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22634–22645, 2023. 2, 6, 7

[12] Yuchao Gu, Xintao Wang, Liangbin Xie, Chao Dong, Gen Li, Ying Shan, and Ming-Ming Cheng. Vqfr: Blind face restoration with vector-quantized dictionary and parallel decoder. In *European Conference on Computer Vision*, pages 126–143. Springer, 2022. 3

[13] Gustav Eje Henter, Simon Alexanderson, and Jonas Beskow. Moglow: Probabilistic and controllable motion synthesis using normalising flows. *ACM Transactions on Graphics (TOG)*, 39(6):1–14, 2020. 3

[14] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE international conference on computer vision*, pages 1501–1510, 2017. 6

[15] Bin Ji, Ye Pan, Yichao Yan, Ruizhao Chen, and Xiaokang Yang. Stylevr: Stylizing character animations with normalizing flows. *IEEE Transactions on Visualization and Computer Graphics*, 2023. 3, 4

[16] 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. In *ACM SIGGRAPH 2022 Conference Proceedings*, pages 1–10, 2022. 2, 3, 6, 7

[17] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. 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*, pages 694–711. Springer, 2016. 6

[18] Tero Karras, Timo Aila, Samuli Laine, Antti Herva, and Jaakko Lehtinen. Audio-driven facial animation by joint end-to-end learning of pose and emotion. *ACM Transactions on Graphics (TOG)*, 36(4):1–12, 2017. 2

[19] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 6

[20] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[21] Durk P Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. *Advances in neural information processing systems*, 31, 2018. 3, 4

[22] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3387–3396, 2022. 2, 3

[23] Bryan Lim and Stefan Zohren. Time-series forecasting with deep learning: a survey. *Philosophical Transactions of the Royal Society A*, 379(2194):20200209, 2021. 2

[24] 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. 5

[25] 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. *arXiv preprint arXiv:2301.01081*, 2023. 2, 3

[26] Debin Meng, Xiaojiang Peng, Kai Wang, and Yu Qiao. Frame attention networks for facial expression recognition in videos. In *2019 IEEE international conference on image processing (ICIP)*, pages 3866–3870. IEEE, 2019. 6

[27] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014. 2

[28] Niranjan D Narvekar and Lina J Karam. A no-reference image blur metric based on the cumulative probability of blur detection (cpbd). *IEEE Transactions on Image Processing*, 20(9):2678–2683, 2011. 6

[29] Evonne Ng, Hanbyul Joo, Liwen Hu, Hao Li, Trevor Darrell, Angjoo Kanazawa, and Shiry Ginosar. Learning to listen: Modeling non-deterministic dyadic facial motion. 2022. 3

[30] AaronVanDen Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. 3

[31] Trevine Oorloff and Yaser Yacoob. Robust one-shot face video re-enactment using hybrid latent spaces of stylegan2. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20947–20957, 2023. 2

[32] Ye Pan, Ruisi Zhang, Shengran Cheng, Shuai Tan, Yu Ding, Kenny Mitchell, and Xubo Yang. Emotional voice puppetry. *IEEE Transactions on Visualization and Computer Graphics*, 29(5):2527–2535, 2023. 2

[33] Ye Pan, Shuai Tan, Shengran Cheng, Qunfen Lin, Zijiao Zeng, and Kenny Mitchell. Expressive talking avatars. *IEEE Transactions on Visualization and Computer Graphics*, 2024. 2

[34] Pat Pataranutaporn, Valdemar Danry, Joanne Leong, Parinya Punpongsanon, Dan Novy, Pattie Maes, and Misha Sra. Ai-generated characters for supporting personalized learning and well-being. *Nature Machine Intelligence*, 3(12):1013–1022, 2021. 1

[35] Ziqiao Peng, Haoyu Wu, Zhenbo Song, Hao Xu, Xiangyu Zhu, Jun He, Hongyan Liu, and Zhaoxin Fan. Emotalk: Speech-driven emotional disentanglement for 3d face animation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20687–20697, 2023. 2

[36] Ziqiao Peng, Haoyu Wu, Zhenbo Song, Hao Xu, Xiangyu Zhu, Hongyan Liu, Jun He, and Zhaoxin Fan. Emotalk: Speech-driven emotional disentanglement for 3d face animation. 2023. 2

[37] 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, 6, 7

[38] Yurui Ren, Ge Li, Yuanqi Chen, Thomas H Li, and Shan Liu. Pirenderer: Controllable portrait image generation via semantic neural rendering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13759–13768, 2021. 6

[39] DaniloJimenez Rezende and Shakir Mohamed. Variational inference with normalizing flows. *International Conference on Machine Learning,International Conference on Machine Learning*, 2015. 2, 3

[40] Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *International conference on machine learning*, pages 1530–1538. PMLR, 2015. 4

[41] Maximilian Seitzer. pytorch-fid: FID Score for PyTorch. `https://github.com/mseitzer/pytorch-fid`, 2020. Version 0.3.0. 6

[42] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 6

[43] Sanjana Sinha, Sandika Biswas, Ravindra Yadav, and Brojeshwar Bhowmick. Emotion-controllable generalized talking face generation. *international joint conference on artificial intelligence*, 2022. 2

[44] Yang Song, Jingwen Zhu, Dawei Li, Andy Wang, and Hairong Qi. Talking face generation by conditional recurrent adversarial network. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, 2019. 2

[45] Stefan Stan, Kazi Injamamul Haque, and Zerrin Yumak. Facediffuser: Speech-driven 3d facial animation synthesis using diffusion. In *ACM SIGGRAPH Conference on Motion, Interaction and Games (MIG '23), November 15–17, 2023, Rennes, France*, New York, NY, USA, 2023. ACM. 1

[46] 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. 1, 2, 6, 7

[47] Shuai Tan, Bin Ji, Yu Ding, and Ye Pan. Say anything with any style. *arXiv preprint arXiv:2403.06363*, 2024. 2

[48] Shuai Tan, Bin Ji, and Ye Pan. Style2talker: High-resolution talking head generation with emotion style and art style. *arXiv preprint arXiv:2403.06365*, 2024. 2

[49] 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. 1

[50] Guanzhong Tian, Yi Yuan, and Yong Liu. Audio2face: Generating speech/face animation from single audio with attention-based bidirectional lstm networks. In *2019 IEEE international conference on Multimedia & Expo Workshops (ICMEW)*, pages 366–371. IEEE, 2019. 1

[51] Konstantinos Vougioukas, Stavros Petridis, and Maja Pantic. Realistic speech-driven facial animation with gans. *International Journal of Computer Vision*, 128(5):1398–1413, 2020. 1

[52] 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. 3, 6, 7

[53] Jiayu Wang, Kang Zhao, Shiwei Zhang, Yingya Zhang, Yujun Shen, Deli Zhao, Jingren Zhou, Alibaba Group, and Ant Group. Lipformer: High-fidelity and generalizable talking face generation with a pre-learned facial codebook. 3

[54] Jiadong Wang, Xinyuan Qian, Malu Zhang, Robby T Tan, and Haizhou Li. 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*, pages 14653–14662, 2023. 1

[55] Jianrong Wang, Yaxin Zhao, Li Liu, Tianyi Xu, Qi Li, and Sen Li. Emotional talking head generation based on memory-sharing and attention-augmented networks. 2023. 2

[56] Jianrong Wang, Yaxin Zhao, Li Liu, Tianyi Xu, Qi Li, and Sen Li. Emotional talking head generation based on memory-sharing and attention-augmented networks. *arXiv preprint arXiv:2306.03594*, 2023. 2

[57] 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 *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXI*, pages 700–717. Springer, 2020. 2, 5, 6, 7, 8

[58] Kaisiyuan Wang, Changcheng Liang, Hang Zhou, Jiaxiang Tang, Qianyi Wu, Dongliang He, Zhibin Hong, Jingtuo Liu, Errui Ding, Ziwei Liu, and Jingdong Wang. Robust video portrait reenactment via personalized representation quantization. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(2):2564–2572, 2023. 3

[59] S Wang, L Li, Y Ding, C Fan, and X Yu. Audio2head: Audio-driven one-shot talking-head generation with natural head motion. In *International Joint Conference on Artificial Intelligence*. IJCAI, 2021. 2, 8

[60] Suzhen Wang, Lincheng Li, Yu Ding, and Xin Yu. One-shot talking face generation from single-speaker audio-visual correlation learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2531–2539, 2022. 2, 8

[61] 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*, 2004. 6

[62] Zhouxia Wang, Jiawei Zhang, Runjian Chen, Wenping Wang, and Ping Luo. Restoreformer: High-quality blind face restoration from undegraded key-value pairs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17512–17521, 2022. 3, 6

[63] Jinbo Xing, Menghan Xia, Yuechen Zhang, Xiaodong Cun, Jue Wang, and Tien-Tsin Wong. Codetalker: Speech-driven 3d facial animation with discrete motion prior. 2023. 3

[64] 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. 2023. 3

[65] 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: Computer Vision and Pattern Recognition*, 2020. 2

[66] Egor Zakharov, Aliaksandra Shysheya, Egor Burkov, and Victor Lempitsky. Few-shot adversarial learning of realistic neural talking head models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9459–9468, 2019. 2

[67] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 6

[68] 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. *arXiv preprint arXiv:2211.12194*, 2022. 8

[69] Zhimeng Zhang, Lincheng Li, Yu Ding, and Changjie Fan. Flow-guided one-shot talking face generation with a high-resolution audio-visual dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3661–3670, 2021. 6, 7

[70] Weizhi Zhong, Chaowei Fang, Yinqi Cai, Pengxu Wei, Gangming Zhao, Liang Lin, and Guanbin Li. Identity-preserving talking face generation with landmark and appearance priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2023. 2, 6, 7

[71] Hang Zhou, Yu Liu, Ziwei Liu, Ping Luo, and Xiaogang Wang. Talking face generation by adversarially disentangled audio-visual representation. *national conference on artificial intelligence*, 2019. 2

[72] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4176–4186, 2021. 6, 7

[73] Shangchen Zhou, Kelvin Chan, Chongyi Li, and Chen Change Loy. Towards robust blind face restoration with codebook lookup transformer. *Advances in Neural Information Processing Systems*, 35:30599–30611, 2022. 3, 6

[74] Yang Zhou, Xintong Han, Eli Shechtman, Jose Echevarria, Evangelos Kalogerakis, and Dingzeyu Li. Makelttalk: speaker-aware talking-head animation. *ACM Transactions on Graphics*, 2020. 2

[75] 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, 8