FG-EmoTalk: Talking Head Video Generation with Fine-Grained Controllable Facial Expressions

Zhaoxu Sun[1], **Yuze Xuan**[1], **Fang Liu**[2*], **Yang Xiang**[1]

[1]Xiaobing.ai

[2]State Key Laboratory of Media Convergence and Communication, Communication University of China

sunzhaoxu@xiaobing.ai, ozawaoy_xyz@bupt.edu.cn, fangliu@cuc.edu.cn, xiangyang@xiaobing.ai

Abstract

Although deep generative models have greatly improved one-shot video-driven talking head generation, few studies address fine-grained controllable facial expression editing, which is crucial for practical applications. Existing methods rely on a fixed set of predefined discrete emotion labels or simply copy expressions from input videos. This is limiting as expressions are complex, and methods using only emotion labels cannot generate fine-grained, accurate or mixed expressions. Generating talking head video with precise expressions is also difficult using 3D model-based approaches, as 3DMM only models facial movements and tends to produce deviations. In this paper, we propose a novel framework enabling fine-grained facial expression editing in talking face generation. Our goal is to achieve expression control by manipulating the intensities of individual facial Action Units (AUs) or groups. First, compared with existing methods which decouple the face into pose and expression, we propose a disentanglement scheme to isolates three components from the human face, namely, appearance, pose, and expression. Second, we propose to use input AUs to control muscle group intensities in the generated face, and integrate the AUs features with the disentangled expression latent code. Finally, we present a self-supervised training strategy with well-designed constraints. Experiments show our method achieves fine-grained expression control, produces high-quality talking head videos and outperforms baseline methods.

Introduction

Talking head generation has attracted much attention in the field of computer vision over recent years. Though tremendous progress has been made in enhancing the visual quality of generated videos, most existing studies aim at producing more realistic videos (Ren et al. 2021; Wang et al. 2022) or focus on the audio-based lip synchronization (Prajwal et al. 2020b; Cheng et al. 2022). Recently, a few facial expression editing in talking head works have been proposed (de Barros Reis, Dornhofer Paro Costa, and De Martino 2020; Li et al. 2021; Liang et al. 2022), which conduct emotional video synthesis at a coarse granularity to generate specific emotions like happy or sad. These methods either transfer expressions from input videos (Cheng et al. 2022) by assigning expressions frame-by-frame from input video templates, or enable expression editing using predefined emotions (Sun et al. 2022). However, fixed emotions can only represent limited types of emotions in a coarse-grained and discrete manner, making it difficult to achieve natural and precise emotion editing. In addition, emotions from driving videos also introduce ambiguity if input emotions are misjudged. Talking head video generation with fine-grained controllable facial expressions [1] is still an unaddressed problem.

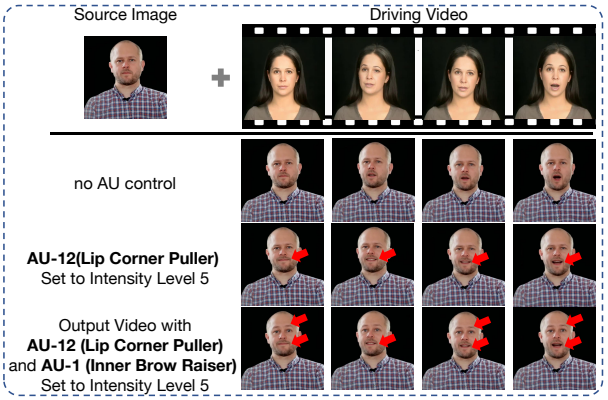

Figure 1: Our FG-EmoTalk enables fine-grained facial expression control via taking input AUs to control the activation intensities of muscle groups in the generated talking face. Our method is subject-agnostic, and can be applied for both audio-driven and video-driven talking head synthesis.

In this work, we address the fine-grained expression editing task in talking head videos, enabling overall emotion control as well as editing expression details with facial AU constraints. Our method uses AUs (Ekman and Friesen 1978) as the fundamental units of expression, and the goal is to control fine-grained expression by manipulating individual AUs or groups. For example, activating AU-12 (lip corner puller) when generating a smile, or manipulating inner brow raiser by setting AU-1 intensity to level 5 (see Fig. 1).

*Fang Liu is the corresponding author.

[1]Note the facial expression here is different from emotion. Emotion is referred to as a specific type of facial expression with a set of predefined AUs' intensities in this paper.

This task is useful yet challenging due to two entangled issues: 1) isolating expression-specific features for more fine-grained granular control, and 2) providing AU-based control and alignment of human faces with expressions. The most related works (Ren et al. 2021; Yin et al. 2022; Pang et al. 2023) adopt feature disentanglement mechanisms to extract pose and expression from input videos. However, we have found these pose and expression features also contain appearance information in our pilot study. Starting from this point, to obtain purer and more efficient expression features while retaining facial appearance, we propose a novel face disentanglement framework isolating not only pose and expression but also appearance-specific features in talking head. Our disentanglement method is able to obtain more efficient appearance-invariant expression features. We also present an AU encoding module, and the learned AU features can be leveraged together with the expression features extracted from the disentanglement scheme to edit latent codes and generate talking heads with fine-grained expression details. Considering the lack of source and generated talking head paired data with fine-grained expression annotations, we design a self-supervised network training strategy, where talking head generated from separated latent codes under different loss constraints are set as supervisions to disentangle appearance, expression, and pose components. Our FG-EmoTalk also enables audio-driven talking head generation. By designing a Wav2Vec2 (Baevski et al. 2020) based module as the audio encoder, the extracted audio features can be fused with the AU features to edit the latent code for talking head generation. In this way, our proposed method can be applied in both audio-driven and video-driven talking head video generation tasks.

The main contributions of our work are three-fold:

- To the best of our knowledge, we first address the *fine-grained controllable* facial expression editing task in talking head video generation. We propose an end-to-end video generation framework that enables expression editing by integrating facial AUs to control muscle group action intensities for expression details.
- We propose a feature decomposition mechanism to disentangle talking head features into three components, namely appearance, pose, and expression. The isolated expression component is then combined with input AUs for fine-grained expressions generation.
- We present a self-supervised network training strategy with well-designed constraints and two new losses, namely appearance loss and expression loss. Extensive experiments demonstrate that our method generates videos with high visual quality and appropriate fine-grained expression details.

Related Work

Talking Head Video Generation

Talking head video generation can be used for a lot of downstream applications such as video conferencing, digital characters, movie special effects, etc. Existing studies can be mainly categorized into two types: audio-driven (Vougioukas, Petridis, and Pantic 2018; Zhou et al. 2020; Lu, Chai, and Cao 2021; Song et al. 2022) and video-driven methods (Ha et al. 2020; Drobyshev et al. 2022). Audio-driven methods aim to maintain audio-lip synchronization of generated animations of human faces, which can be further classified into Generative Adversarial Network (GAN) based methods and 3D Morphable Models (3DMMs) based methods. Most GAN-based methods only generate images depicting the mouth-related region (Prajwal et al. 2020a; Yin et al. 2022; Zhou et al. 2021), and 3D-based approaches usually extract Mel-Frequency Cepstral Coefficient (MFCC) features from the input audio to estimate 3DMM expression coefficients (Zhang et al. 2022) or vertices offsets (Fan et al. 2022). Compared to 3DMM approaches which only model facial movements, our method also incorporates head and shoulder movements. Additionally, 3DMM approaches tend to produce deviations, and smoothing strategies like in PIRender (Ren et al. 2021) would introduce expression inaccuracies.

Compared with audio-driven methods, video-driven methods can utilize richer information contained in the input video to generate more natural and realistic results, which can be roughly classified into 2D keypoint-based methods (Siarohin et al. 2019; Zhao and Zhang 2022), 2D GAN-based methods (Wang et al. 2022; Yin et al. 2022), and 3D-model-based networks (Lahiri et al. 2021; Hong et al. 2022; Ma et al. 2023). 2D keypoint-based methods first compute the transfer matrix via matching keypoint pairs between the source image and the driving image, then wrap the source image to get the dense flow, and finally generate images with a GAN generator (Hong et al. 2022). 2D GAN-based methods mostly utilize the prior information obtained by StyleGAN (Karras et al. 2020) and manipulate the talking head video generation by conducting metric learning or conditional injection (Zhou et al. 2021; Tzaban et al. 2022). 3D-based networks leverage 3DMM coefficients for human face reconstruction, but the generated videos tend to have facial inconsistency problems due to the inaccurate recognition of 3DMM coefficients (Ren et al. 2022).

Our method uses a driving video and a source image as the input to generate talking head with high quality and ideal facial expression details. Moreover, our framework can be further extended to audio-based talking head video generation.

Emotion Editing in Talking Head Videos

Most talking head video generation research concentrate on enhancing the visual quality of the output video, while the facial emotions in videos are neglected (Zhou et al. 2021). Very recently, a few studies generates talking head with facial expressions according to whole input video (Cheng et al. 2022), where the produced emotions usually lack stability. Besides, these methods can only transfer expression in the driving video and do not support fine-grained expression control (de Barros Reis, Dornhofer Paro Costa, and De Martino 2020; Liang et al. 2022). For example, EAMM (Ji et al. 2022) aims at generating one-shot emotional talking faces on arbitrary subjects, and it extract emotion patterns from the source video. Emotalk (Peng et al. 2023) is a speech-driven 3D face animation method, while our approach can be applied in both video-driven and audio-driven. GC-AVT

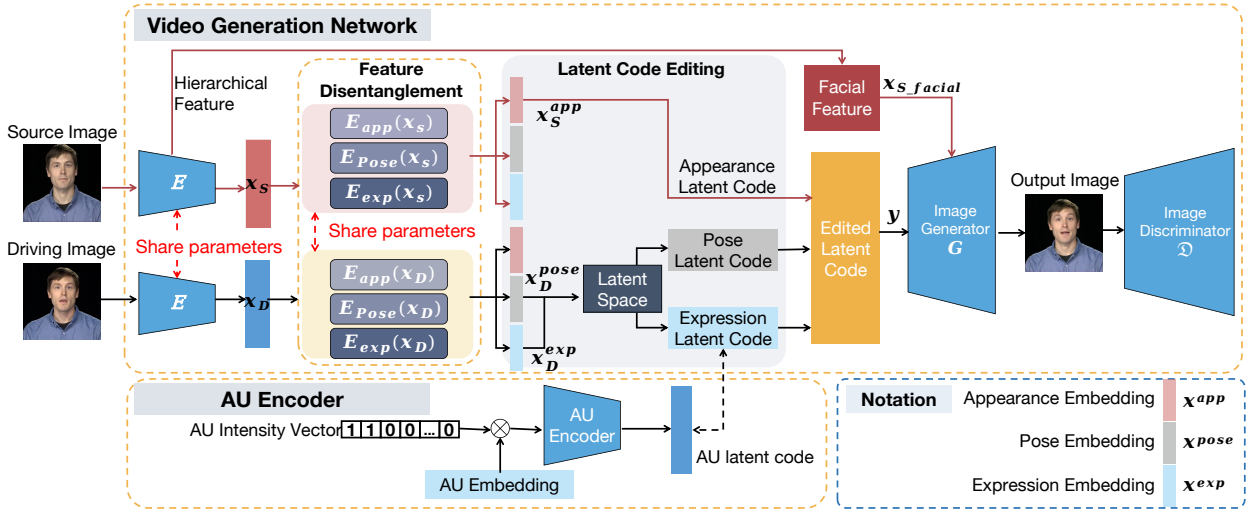

Figure 2: The framework of our FG-EmoTalk. Our method mainly contains a video generation network and an AU encoder module. Given an input source image and a driving image, image features are first extracted with the image encoder, and then disentangled into three components, *i.e.,* the appearance embedding, the pose embedding and the expression embedding. Then, the pose and expression embeddings are fused in a latent space. The expression embedding is further constrained by the AU feature. Finally, the edited latent code is combined with the facial feature to generate the talking head via an image decoder.

(Liang et al. 2022) disassembles the driving image into a cropped mouth, a masked head and a upper face, to implement expressive generation. Fine-grained emotion editing in talking head generation remains an unsolved task. Different from the above methods which extracts facial expression dynamics and motion features, we propose to disentangle not only expression and motion, but also appearance information. Recently, very few studies (Chen et al. 2021, 2022) leverage AUs (Zhang et al. 2021a) to help the talking head video generation. However, these studies focus on enhancing image quality and lip-sync accuracy, while this paper uses AUs as the input for fine-grained expression control.

Method

The architecture of our proposed network is illustrated in Fig. 2, which supports one-shot talking head video generation with a source image, a driving video and input intensities of AUs. Our talking head generation framework is built on a GAN-based structure (Wang et al. 2022), and mainly consists of a video generation network and an AU encoder. There are two key modules in the video generation network: (i) the feature disentanglement module which learns the appearance, pose and expression embedding from the image features, and (ii) the latent code editing module integrates the expression embedding with input AU features and construct the latent code for final image generation. The AU encoder module extracts the AU latent code of the input AU intensity vector for target expression editing.

Video Generation Network

The video generation network mainly consists of an image encoder, a feature disentanglement module that contains an appearance extractor, a pose extractor and an expression extractor, a latent code editing module, an image generator, and an image discriminator. Given an input source image S and a frame from the driving video (denoted as driving image) D, we first extract image features from them with a CNN-based image encoder and disentangle the image features into three components, namely the appearance embedding, the pose embedding and the expression embedding. Then, we follow (Wang et al. 2022) to fuse the pose embedding and the expression embedding which represents the motion information with an orthogonal dictionary in a latent space. The expression embedding is further constrained by the AU feature of the AU encoder module. Finally, the edited latent code composed of appearance, pose, and expression information is combined with the hierarchical facial feature to generate the output talking head via an image generator.

Denoting the image encoder as $\mathbf{E}(\cdot)$, we formulate the pipeline of the video generation network as:

$$x_S = \mathbf{E}(S), x_D = \mathbf{E}(D) \tag{1}$$

where x_S and x_D represent the output source and driving image feature. Besides, we also extract the hierarchical facial features of the source image S as x_{S_facial} with $\mathbf{E}(\cdot)$, which will be used in the latter image generation process.

Then, three feature extractors are designed to disentangle the source and driving image features, *i.e.,* an appearance extractor E_{app}, an expression extractor E_{exp}, and a pose extractor E_{pose}. The appearance extractor learns image facial appearance, such as the shape of the face and the size of facial features. We set a specific appearance extractor to ensure that the outputs of the pose and expression extractors contain only motion features and not facial appearance information. The appearance extractor E_{app} consists of a fully connected neural network that employs self-attention mechanisms. These three extractors share the same structure as

in (Wang et al. 2022) and the same parameters, which helps guarantee that the latent representations encoding facial appearance, expression, and body posture are embedded in the same latent space. The disentanglement process and the latent code editing process can be formulated as:

$$x_n^m = E_m(x_n), m \in \{app, pose, exp\}, n \in \{S, D\} \quad (2)$$

where x_n^m represents the appearance, pose or expression embedding of the source or driving image. Since our FG-EmoTalk aims to keep appearance inconsistency from the source image and be similar to the driving image in pose and expression, we use the appearance embedding of the source x_S^{app}, and the pose and expression embeddings of the driving image (x_D^{pose} and x_D^{exp}), for latter image generation process.

Furthermore, the motion-related embeddings of the driving image, *i.e.,* x_D^{pose} and x_D^{exp}, are fused with an orthogonal dictionary in a latent space $Dic(\cdot)$, which is further added with the appearance embedding of the source image x_S^{app} to form the final edited latent code y:

$$y = x_S^{app} + Dic(x_D^{exp}, x_D^{pose}) \quad (3)$$

We finally feed the edited latent code y and the hierarchical facial features x_{S_facial} into the generator $\mathbf{G}(\cdot)$ and obtain the generated output image S':

$$S' = \mathbf{G}(x_{S\_facial}, y) \quad (4)$$

AU Encoder Module

To achieve fine-grained control of facial expressions, we design an AU encoder module (see the lower left corner of Fig. 2) to incorporate with the expression latent code for facial expression details control. The AU Encoder consists of a learnable *AU Embedding* matrix and a Gated-GCN network.

The input of the AU encoder module is an *AU intensity vector* representing the type and intensity of each action unit. The AU intensity vector is first multiplied by the corresponding *AU Embedding*, and the output is encoded with the Gated-GCN, where an averaging graph pooling layer is applied to extract graph features. The reason for using Gated-GCN to encode AU intensity features is to propagate features based on the relationship between AUs as in (Luo et al. 2022). During training, the output AU features of the AU encoder module are set as supervision for the expression latent code learning in the video generation network. During testing, the expression latent can be added or directly replaced with the AU embedding. In Fig. 1, row 3/4 demonstrates generation by adding the expression latent with the AU embedding, allowing specific AUs to change while preserving other expressions. Experiments demonstrate the effectiveness of editing facial expression details by modifying the AU intensity vector (refer to the Experiments section).

Self-Supervised Training Strategy

Due to the lack of paired talking head source and generated data with fine-grained facial expression details, and there is no dataset involving separated facial expressions and pose movements for the same talking head either, we design a self-supervised learning strategy to train our FG-EmoTalk.

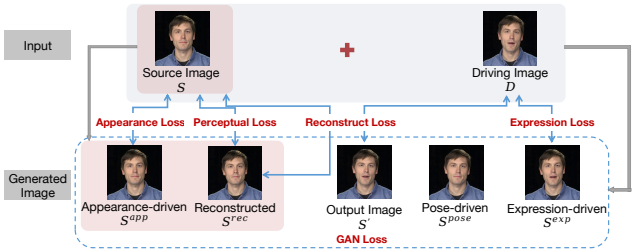

Figure 3: The loss calculation process.

Images generated with separated latent codes using different loss constraints are leveraged to conduct the disentanglement of appearance, expression, and pose information.

Specifically, taking the source image S and the driving image D as input, our network generates the image $S' = \mathbf{G}(x_{S\_facial}, y)$ as described in Equation 4, where y is the edited latent code composed of the appearance of the source image x_S^{app}, and the fused expression and pose feature of the driving image $Dic(x_D^{exp}, x_D^{pose})$. Besides the target image S', the other four images conditioned on the edited latent code with slight changes are also generated during the training process. We use different combinations of latent code y with the facial feature of the input source image x_{S_facial} to generate the images: (1) the appearance-driven image S^{app} generated only with the appearance latent code; (2) the expression-driven image S^{exp} generated with the appearance latent code and the expression latent code; (3) the pose-driven image S^{pose} generated with the appearance latent code and the pose latent code, and (4) the reconstructed images S^{rec} generated only with the combination of its all three latent codes. We formulate the procedure as:

$$\begin{aligned} S^{app} &= \mathbf{G}(x_{S\_facial}, E_{app}(x_S)) \\ S^{exp} &= \mathbf{G}(x_{S\_facial}, E_{app}(x_S) + Dic(E_{exp}(x_D))) \\ S^{pose} &= \mathbf{G}(x_{S\_facial}, E_{app}(x_S) + Dic(E_{pose}(x_D))) \\ S^{rec} &= \mathbf{G}(x_{S\_facial}, E_{app}(x_S) + Dic(E_{exp}(x_S)) \\ &\quad + Dic(E_{pose}(x_S))) \end{aligned} \quad (5)$$

To effectively disentangle the appearance, expression, and pose information, we design a set of constraints on the above intermediate output images via loss functions. Specifically, four types of image pairs are used as the source and driving image during the network training process to ensure the appearance ($\langle S^{app}, S\rangle$), expression ($\langle S^{exp}, D\rangle$), and content ($\langle S', D\rangle$, $\langle S^{rec}, S\rangle$) consistency. These constraints offer effective supervision for the disentanglement of appearance, expression, and pose latent codes. Fig. 3 shows the conditional generated images and the self-supervised training procedure. See more analysis in our experiments in Sec. .

Loss Function

Our loss function consists of five loss terms.

Appearance Loss $\mathcal{L}_{app}$. The appearance loss constrains the generated talking head is the same identity as the input image. Popular appearance loss computing methods use arcface (Deng et al. 2019) as the backbone to measure the

similarity of the input human face and the generated one. However, talking head generation in our work involves not only the human face but also the head and shoulder regions. Inspired by the idea in the ReID field, we adopt a pre-trained backbone in (Zheng et al. 2019) (denoted as $\mathcal{L}_A(\cdot)$) to compute the appearance loss $\mathcal{L}_{app}$:

$$\mathcal{L}_{\text{app}} = \mathcal{L}_A\left(S^{app}, S\right) \tag{6}$$

Expression Loss $\mathcal{L}_{exp}$. Previous methods have used facial expression recognition network (Pang et al. 2023) to calculate the expression loss. However, on one hand, the performance of facial expression recognition is instability. On the other hand, the widely used datasets (Zhang et al. 2021b; Zhu et al. 2022) in the talking head field do not have enough data with significant facial expressions. The existing method cannot well evaluate the quality of the expression latent code. To enable talking heads generation with a greater degree of expression variation, we use the MicronBERT (Nguyen et al. 2023) (denoted as $\mathcal{L}_E$) to compute the distance of facial expressions:

$$\mathcal{L}_{\text{exp}} = \mathcal{L}_E\left(S^{exp}, D\right) \tag{7}$$

Perceptual loss $\mathcal{L}_{per}$. To enhance the vividness of the generated results, we adopt a perceptual loss based on VGG-16 (Johnson, Alahi, and Fei-Fei 2016) (denoted as $\mathcal{L}_P$), to constrain the similarity of the following two pairs of images:

$$\mathcal{L}_{\text{per}} = \mathcal{L}_P\left(S', D\right) + \mathcal{L}_P\left(S^{rec}, S\right). \tag{8}$$

Reconstruction loss $\mathcal{L}_{rec}$. We use Mean Absolute Error (MAE) loss (denoted as $\mathcal{L}_R$) to calculate the reconstruction error between image pairs:

$$\mathcal{L}_{\text{rec}} = \mathcal{L}_R\left(S', D\right). \tag{9}$$

GAN loss $\mathcal{L}_{gan}$. We use a discriminator to distinguish generated images from source images, and use a non-saturated adversarial loss $\mathcal{L}_G$ to measure the error of the generated results:

$$\begin{aligned}\mathcal{L}_{\text{gan}} &= \mathcal{L}_G(S') + \mathcal{L}_G(S^{app}) + \mathcal{L}_G(S^{exp}) \\ &\quad + \mathcal{L}_G(S^{pose}) + \mathcal{L}_G(S^{rec})\end{aligned} \tag{10}$$

Finally the overall loss function is a combination of the five loss terms:

$$\mathcal{L} = \lambda_{app}\mathcal{L}_{\text{app}} + \lambda_{exp}\mathcal{L}_{\text{exp}} + \lambda_{per}\mathcal{L}_{\text{per}} + \mathcal{L}_{\text{rec}} + \mathcal{L}_{\text{gan}} \tag{11}$$

where λ_{app}, λ_{exp} and λ_{per} are hyper-parameters used in network training process.

Experiments

Datasets. We use the HDTF (Zhang et al. 2021b) and CelebV-HQ (Zhu et al. 2022) datasets which have no emotion or AU annotation to train our video generation network (see the upper part of Fig. 2). As for the evaluation, we select 2,000 videos from the HDTF dataset that did not appear in the training set. Moreover, the MEAD dataset (Wang et al. 2020) contains emotional talking face videos of different actors speaking with 8 emotion categories. We also randomly select 2,000 videos from the MEAD dataset for testing to validate the cross-dataset generalization ability of our FG-EmoTalk. All videos are resized to a resolution of 512x512. We did not geometrically align faces, instead allowing free motion within a fixed bounding box.

Method	$ACC_{emo}\uparrow$	PSNR↑	SSIM ↑
EAMM	0.5435	30.0120	0.8577
VideoRetalking	0.5773	31.2131	0.8437
Ours	**0.6102**	**31.4556**	**0.8802**

Table 1: Quantitative evaluation on facial expression quality.

We used the DISFA dataset (Mavadati et al. 2013) to train our AU Encoder Module for AU-based expression editing. Due to annotation issues, we filtered some poorly labeled examples and AUs affecting mouth shapes. We retained 7 AUs for expression control, shown in Fig. 5.

Evaluation Metrics. We evaluate talking head generation methods across three factors: (1) facial expression quality: we adopt the accuracy of emotion classification (ACC_{emo}), PSNR, and SSIM to measure the quality of the expression in the generated videos; (2) visual quality: we use learned perceptual image patch similarity (LPIPS), peak signal-to-noise ratio (PSNR), structural similarity index (SSIM), average pose coefficient distance (APD), and average emotion coefficient distance (AED) to evaluate the visual quality of the generated videos; (3) disentanglement of pose and expression: we use mean absolute error (MAE) to measure the accuracy of predicted pose and expression features. In all the tables, ↓ indicates "the smaller the better", and ↑ indicates "the larger the better".

Implementation Details. We implemented our framework in Pytorch. During training, we first train the video generation network and then train the AU encoder with the parameters of the video generation network fixed to enable fine-grained expression editing.During testing, the expression latent can be either added or directly replaced with the AU embedding. In Fig. 1, row 3/4 show the generation results by adding the expression latent with the AU embedding, allowing specific AUs to change while preserving other expressions. All experiments were conducted with 4 NVIDIA Tesla A10 GPUs. We used the Adam optimizer with a learning rate of 0.002. The hyperparameters λ_{app}, λ_{exp}, and λ_{per} were set to 100.0, 100.0, and 10.0 respectively in the training stage. To improve teeth clarity in generated images, we adopted fine-tuning the GFP-GAN (Wang et al. 2021) as an optional post-processing step. By fine-tuning for 100 epochs on CelebV-HQ, the quality of teeth in generated talking head has been significantly improved.

Compare with State-Of-The-Art Methods

Facial Expression Quality. Facial expression control using AU intensity is the focus of our FG-EmoTalk. We investigate the expression editing performance of our method with single AU and Muliple AUs.

Single AU. Talking head generation results with each AU are shown in Fig. 5, demonstrating that our method can edit

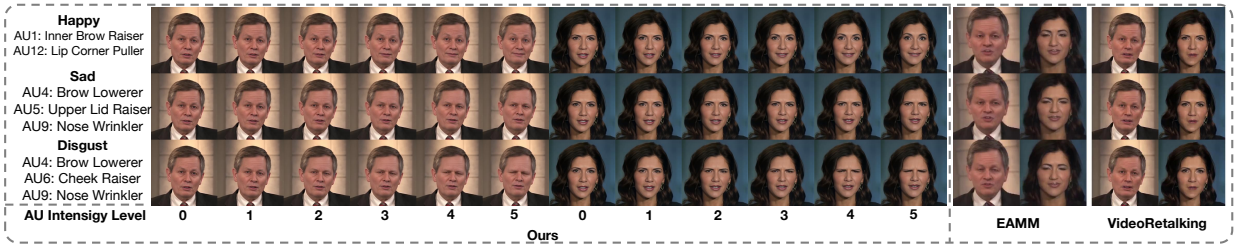

Figure 4: Talking head generation example results with emotions involving combinations of multiple AUs. Left: results of our method with different AUs' intensities. Right: results of EAMM and VideoRetalking with emotion labels.

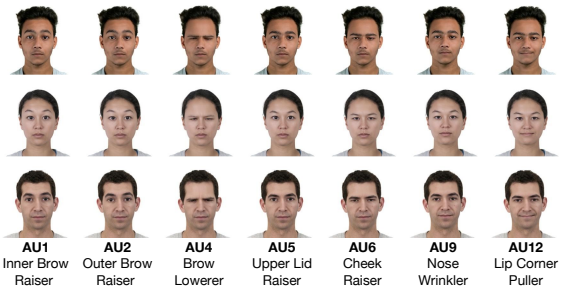

Figure 5: Talking head generation results with single AU. Input images are from CFD dataset (Lakshmi et al. 2021).

Figure 6: Visual quality comparison with SOTA methods. Top two rows: same-identity reenactment. Bottom two rows: cross-identity reenactment.

Method	TPSM	StyleHeat	LIA	DPE	Ours
LPIPS ↓	0.1587	0.3374	**0.1563**	0.1598	0.1565
PSNR ↑	31.1108	20.6398	31.3126	30.2071	**31.4725**
SSIM ↑	0.7928	0.5674	0.7834	0.7534	**0.7970**
APD ↓	0.0358	0.0200	0.0301	0.0089	**0.0067**
AED ↓	0.0075	0.0152	0.0174	0.0153	**0.0053**
APD ↓	0.0199	0.0232	0.0139	0.0093	**0.0088**
AED ↓	0.0286	0.0632	0.0283	0.0528	**0.0185**

Table 2: Quantitative comparisons on visual quality. Rows (2-6): same-identity; Rows (7-8): cross-identity.

fine-grained expressions by integrating AUs to control muscle group intensities.

Multiple AUs. To eliminate the impact of the original expressions of driving images, we further choose to conduct an audio-driven talking head generation experiment to explore the performance of expression control with multiple AUs. We compare two representative emotional talking head generation methods, *i.e.,* EAMM (Ji et al. 2022) and VideoRetalking (Cheng et al. 2022), which generate talking head with emotion labels (not fine-grained). We first follow (Ling et al. 2020) to establish the correspondence between three emotions and their corresponding combinations of AUs (happy: AU1+AU12, sad: AU4+AU5+AU9, disgust: AU4+AU6+AU9). Then we randomly select 100 groups of images and audio combinations from the test set of CelebV-HQ for evaluation. In order to eliminate the influence of mouth shapes generated by these methods on the evaluation metrics, we adopt an emotion recognition network (Amos, Ludwiczuk, and Satyanarayanan 2016) to measure the quality of generated expressions. We also include common PSNR and SSIM indicators as the metrics. The results are shown in Tab. 1, indicating that our method achieves the best performance. From the visual results shown in Fig. 4, we can observe that the face reconstruction performance of EAMM is not satisfactory, and the emotional expression editing results of our method are better than those of VideoRetalking. In addition, we also show generation results with different AUs' intensities in the left part of Fig. 4, demonstrating that our FG-EmoTalk can edit expressions with variable intensities. The input images are from Voxceleb2 dataset (Chung, Nagrani, and Zisserman 2018), which are not used for training and can be seen as in-the-wild inputs.

Visual Quality Comparison. We compare our method with four state-of-the-art (SOTA) approaches on the visual quality of generated talking heads, *i.e.,* TPSM (Zhao and Zhang 2022), StyleHeat (Yin et al. 2022), LIA (Wang et al. 2022) and DPE (Pang et al. 2023). TPSM is a typical 2D keypoint-based method surpassing First-Order Motion Model. LIA is a typical StyleGAN (Karras et al. 2020)-based method. StyleHeat and DPE are representative works focusing on disentanglement techniques in talking head generation. As shown in Tab. 2, our approach achieves the highest scores across all metrics on both same-identity and cross-identity experimental settings compared to the baselines. Fig. 6 also shows our visual results are superior to these methods in terms of visual quality.

User study. We further conduct a user study to compare our method with the four SOTA methods. We randomly select 10 generation examples from the HDTF dataset using these methods. The participants were presented with the in-

Method	(a)	(b)	(c)	(d)	(e)
TPSM	9.5%	8.0%	11.5%	8.5%	8.5%
StyleHeat	7.0%	6.5%	11.5%	9.0%	8.5%
LIA	14.5%	13.0%	9.0%	12.5%	10.5%
DPE	4.0%	4.5%	4.5%	6.0%	4.5%
Ours	**65.0%**	**68.0%**	**63.5%**	**64.0%**	**68.0%**

Table 3: User study. (a): Naturalness; (b): Appearance; (c): Lip sync; (d): Expression; (e): Overall. The percentages of five methods being selected as the best for corresponding criterion are shown.

Method	SSIM ↑	APD ↓	AED ↓
StyleHeat	0.6535	0.0184	0.0300
DPE	0.8535	0.0135	0.0265
Ours	**0.8595**	**0.0115**	**0.0222**

Table 4: Comparison with SOTA methods on disentanglement efficiency of *pose* and *expression* latent code.

put source photo, the driving video, as well as the five generated videos, arranged in a side-by-side manner to ensure a randomized order. Subsequently, they were tasked with responding to questions of five criterion: naturalness, appearance, lip sync, expression and overall. A total of 20 participants participated in this user study. The results of each method being selected as the best under the criterion in each question are shown in Tab. 3. Our proposed method achieves the best results across all evaluation criteria.

Disentanglement of Pose and Expression. To evaluate the effectiveness of our disentanglement mechanism, we compare our approach with two SOTA disentanglement-based talking head generation methods, *i.e.,* StyleHeat (Yin et al. 2022) and DPE (Pang et al. 2023). As shown in Tab. 4, our method achieves the best results. Fig. 7 shows the visual results generated with only pose or expression latent code, respectively. These results demonstrate that our method performs better than the baselines in imitating the pose and expressions of driving images.

Ablative Study

We further explore the contributions of the loss functions. Specifically, three variants of our model are implemented by removing the appearance loss, the expression loss, or both from our full model. The quantitative results are shown in Tab. 5. We can conclude that: (1) the LPIPS changes significantly without appearance loss, which indicates appearance loss is crucial for capturing visual similarity. (2) The APD and AED scores increase obviously without expression loss, demonstrating that the expression loss is important to disentangle appearance and expression. (3) Removing both appearance and expression losses result in invalid generated images and significantly reduces reconstruction accuracy and expression similarity.

Driven by Audio

Besides the video-driven talking head generation, our framework also enables using audio for controlling the generation.

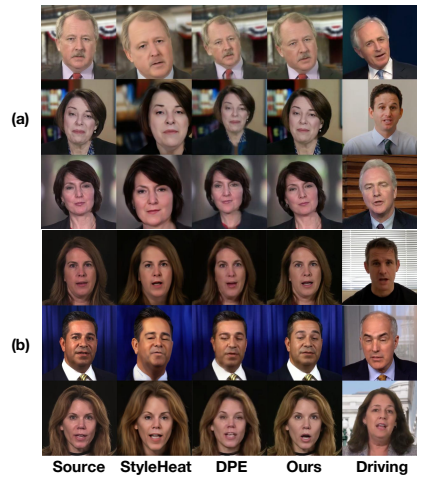

Figure 7: (a) Talking head video examples generated with only *pose* latent code. (b) Talking head video examples generated with only *expression* latent code.

	LPIPS ↓	PSNR ↑	SSIM ↑	APD ↓	AED ↓
(a)	0.1735	29.4573	0.7404	0.0186	0.0214
(b)	0.1690	30.8425	0.7574	0.0307	0.0117
(c)	0.2111	27.3525	0.7092	0.0302	0.0122
Ours	**0.1565**	**31.4725**	**0.7970**	**0.0067**	**0.0053**

Table 5: Quantitative results of ablation study. (a): without $\mathcal{L}_{\text{app}}$; (b): without $\mathcal{L}_{\text{exp}}$; (c): without $\mathcal{L}_{\text{app}}$&$\mathcal{L}_{\text{exp}}$.

Existing audio-driven studies usually control mouth movement with 3DMM coefficients, while we directly align audio features with the disentangled expression latent code to control mouth shape changes. Specifically, we design an audio encoder based on Wav2Vec (Baevski et al. 2020) and ResBlocks (He et al. 2016) to extract audio features, and adopt InfoNCE (Oord, Li, and Vinyals 2018) to constrain the alignment between the output of the audio encoder and the expression latent code of the video generation network. In this way, we can control mouth-shape movements in talking head generation with input audio.

Conclusion

In this paper, we propose a talking head generation framework enabling fine-grained facial expression control. An self-supervised disentanglement strategy is designed to disentangle facial appearance, expression, and posture information. By aligning the expression latent code with AU embedding, our FG-EmoTalk enables fine-grained expression editing using AUs. Extensive experiments show our method is better than baseline methods and can control fine-grained expressions. For future work, on one hand, we will collect new datasets with more fine-grained expression annotations for more precision talking head generation. On the other hand, we will cope up with more interactive modalities to enable more user-friendly talking head generation.

References

Amos, B.; Ludwiczuk, B.; and Satyanarayanan, M. 2016. OpenFace: A general-purpose face recognition library with mobile applications. Technical report, CMU-CS-16-118, CMU School of Computer Science.

Baevski, A.; Zhou, Y.; Mohamed, A.; and Auli, M. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33: 12449–12460.

Chen, S.; Liu, Z.; Liu, J.; and Wang, L. 2022. Talking Head Generation Driven by Speech-Related Facial Action Units and Audio-Based on Multimodal Representation Fusion. *arXiv preprint arXiv:2204.12756*.

Chen, S.; Liu, Z.; Liu, J.; Yan, Z.; and Wang, L. 2021. Talking head generation with audio and speech related facial action units. *arXiv preprint arXiv:2110.09951*.

Cheng, K.; Cun, X.; Zhang, Y.; Xia, M.; Yin, F.; Zhu, M.; Wang, X.; Wang, J.; and Wang, N. 2022. VideoReTalking: Audio-based Lip Synchronization for Talking Head Video Editing In the Wild. In *SIGGRAPH Asia 2022 Conference Papers*, 1–9.

Chung, J. S.; Nagrani, A.; and Zisserman, A. 2018. Voxceleb2: Deep speaker recognition. *arXiv preprint arXiv:1806.05622*.

de Barros Reis, F. A.; Dornhofer Paro Costa, P.; and De Martino, J. M. 2020. Deeply Emotional Talking Head: A Generative Adversarial Network Approach to Expressive Speech Synthesis with Emotion Control. In *ACM SIGGRAPH 2020 Posters*, 1–2.

Deng, J.; Guo, J.; Xue, N.; and Zafeiriou, S. 2019. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4690–4699.

Drobyshev, N.; Chelishev, J.; Khakhulin, T.; Ivakhnenko, A.; Lempitsky, V.; and Zakharov, E. 2022. Megaportraits: One-shot megapixel neural head avatars. In *Proceedings of the 30th ACM International Conference on Multimedia*, 2663–2671.

Ekman, P.; and Friesen, W. V. 1978. Facial action coding system. *Environmental Psychology & Nonverbal Behavior*.

Fan, Y.; Lin, Z.; Saito, J.; Wang, W.; and Komura, T. 2022. Faceformer: Speech-driven 3d facial animation with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18770–18780.

Ha, S.; Kersner, M.; Kim, B.; Seo, S.; and Kim, D. 2020. Marionette: Few-shot face reenactment preserving identity of unseen targets. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, 10893–10900.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Hong, F.-T.; Zhang, L.; Shen, L.; and Xu, D. 2022. Depth-aware generative adversarial network for talking head video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3397–3406.

Ji, X.; Zhou, H.; Wang, K.; Wu, Q.; Wu, W.; Xu, F.; and Cao, X. 2022. Eamm: One-shot emotional talking face via audio-based emotion-aware motion model. In *ACM SIGGRAPH 2022 Conference Proceedings*, 1–10.

Johnson, J.; Alahi, A.; and Fei-Fei, L. 2016. Perceptual Losses for Real-Time Style Transfer and Super-Resolution. arXiv:1603.08155.

Karras, T.; Laine, S.; Aittala, M.; Hellsten, J.; Lehtinen, J.; and Aila, T. 2020. Analyzing and Improving the Image Quality of StyleGAN. In *Proc. CVPR*.

Lahiri, A.; Kwatra, V.; Frueh, C.; Lewis, J.; and Bregler, C. 2021. Lipsync3d: Data-efficient learning of personalized 3d talking faces from video using pose and lighting normalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2755–2764.

Lakshmi, A.; Wittenbrink, B.; Correll, J.; and Ma, D. S. 2021. The India Face Set: International and cultural boundaries impact face impressions and perceptions of category membership. *Frontiers in psychology*, 12: 627678.

Li, L.; Wang, S.; Zhang, Z.; Ding, Y.; Zheng, Y.; Yu, X.; and Fan, C. 2021. Write-a-speaker: Text-based emotional and rhythmic talking-head generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 1911–1920.

Liang, B.; Pan, Y.; Guo, Z.; Zhou, H.; Hong, Z.; Han, X.; Han, J.; Liu, J.; Ding, E.; and Wang, J. 2022. Expressive talking head generation with granular audio-visual control. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3387–3396.

Ling, J.; Xue, H.; Song, L.; Yang, S.; Xie, R.; and Gu, X. 2020. Toward fine-grained facial expression manipulation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVIII 16*, 37–53. Springer.

Lu, Y.; Chai, J.; and Cao, X. 2021. Live speech portraits: real-time photorealistic talking-head animation. *ACM Transactions on Graphics (TOG)*, 40(6): 1–17.

Luo, C.; Song, S.; Xie, W.; Shen, L.; and Gunes, H. 2022. Learning multi-dimensional edge feature-based au relation graph for facial action unit recognition. *arXiv preprint arXiv:2205.01782*.

Ma, Y.; Wang, S.; Hu, Z.; Fan, C.; Lv, T.; Ding, Y.; Deng, Z.; and Yu, X. 2023. Styletalk: One-shot talking head generation with controllable speaking styles. *arXiv preprint arXiv:2301.01081*.

Mavadati, S. M.; Mahoor, M. H.; Bartlett, K.; Trinh, P.; and Cohn, J. F. 2013. Disfa: A spontaneous facial action intensity database. *IEEE Transactions on Affective Computing*, 4(2): 151–160.

Nguyen, X.-B.; Duong, C. N.; Xin, L.; Susan, G.; Han-Seok, S.; and Luu, K. 2023. Micron-BERT: BERT-based Facial Micro-Expression Recognition. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Oord, A. v. d.; Li, Y.; and Vinyals, O. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Pang, Y.; Zhang, Y.; Quan, W.; Fan, Y.; Cun, X.; Shan, Y.; and Yan, D.-m. 2023. DPE: Disentanglement of Pose and Expression for General Video Portrait Editing. *arXiv preprint arXiv:2301.06281*.

Peng, Z.; Wu, H.; Song, Z.; Xu, H.; Zhu, X.; Liu, H.; He, J.; and Fan, Z. 2023. EmoTalk: Speech-driven emotional disentanglement for 3D face animation. *arXiv preprint arXiv:2303.11089*.

Prajwal, K.; Mukhopadhyay, R.; Namboodiri, V. P.; and Jawahar, C. 2020a. 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*, 484–492.

Prajwal, K. R.; Mukhopadhyay, R.; Namboodiri, V. P.; and Jawahar, C. 2020b. 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*, MM '20, 484–492. New York, NY, USA: Association for Computing Machinery. ISBN 9781450379885.

Ren, Y.; Li, G.; Chen, Y.; Li, T. H.; and Liu, S. 2021. Pirenderer: Controllable portrait image generation via semantic neural rendering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 13759–13768.

Ren, Y.; Li, G.; Chen, Y.; Li, T. H.; and Liu, S. 2022. PIRenderer: Controllable Portrait Image Generation via Semantic Neural Rendering. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*.

Siarohin, A.; Lathuilière, S.; Tulyakov, S.; Ricci, E.; and Sebe, N. 2019. First order motion model for image animation. *Advances in Neural Information Processing Systems*, 32.

Song, L.; Wu, W.; Qian, C.; He, R.; and Loy, C. C. 2022. Everybody's talkin': Let me talk as you want. *IEEE Transactions on Information Forensics and Security*, 17: 585–598.

Sun, Z.; Wen, Y.-H.; Lv, T.; Sun, Y.; Zhang, Z.; Wang, Y.; and Liu, Y.-J. 2022. Continuously Controllable Facial Expression Editing in Talking Face Videos. *arXiv preprint arXiv:2209.08289*.

Tzaban, R.; Mokady, R.; Gal, R.; Bermano, A.; and Cohen-Or, D. 2022. Stitch It in Time: GAN-Based Facial Editing of Real Videos. In *SIGGRAPH Asia 2022 Conference Papers*, SA '22. New York, NY, USA: Association for Computing Machinery. ISBN 9781450394703.

Vougioukas, K.; Petridis, S.; and Pantic, M. 2018. End-to-end speech-driven facial animation with temporal gans. *arXiv preprint arXiv:1805.09313*.

Wang, K.; Wu, Q.; Song, L.; Yang, Z.; Wu, W.; Qian, C.; He, R.; Qiao, Y.; and Loy, C. C. 2020. MEAD: A Large-scale Audio-visual Dataset for Emotional Talking-face Generation. In *ECCV*.

Wang, X.; Li, Y.; Zhang, H.; and Shan, Y. 2021. Towards real-world blind face restoration with generative facial prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9168–9178.

Wang, Y.; Yang, D.; Bremond, F.; and Dantcheva, A. 2022. Latent image animator: Learning to animate images via latent space navigation. *arXiv preprint arXiv:2203.09043*.

Yin, F.; Zhang, Y.; Cun, X.; Cao, M.; Fan, Y.; Wang, X.; Bai, Q.; Wu, B.; Wang, J.; and Yang, Y. 2022. StyleHEAT: One-shot high-resolution editable talking face generation via pre-trained StyleGAN. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XVII*, 85–101. Springer.

Zhang, C.; Zhao, Y.; Huang, Y.; Zeng, M.; Ni, S.; Budagavi, M.; and Guo, X. 2021a. Facial: Synthesizing dynamic talking face with implicit attribute learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, 3867–3876.

Zhang, W.; Cun, X.; Wang, X.; Zhang, Y.; Shen, X.; Guo, Y.; Shan, Y.; and Wang, F. 2022. SadTalker: Learning Realistic 3D Motion Coefficients for Stylized Audio-Driven Single Image Talking Face Animation. *arXiv preprint arXiv:2211.12194*.

Zhang, Z.; Li, L.; Ding, Y.; and Fan, C. 2021b. 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*, 3661–3670.

Zhao, J.; and Zhang, H. 2022. Thin-plate spline motion model for image animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3657–3666.

Zheng, Z.; Yang, X.; Yu, Z.; Zheng, L.; Yang, Y.; and Kautz, J. 2019. Joint discriminative and generative learning for person re-identification. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zhou, H.; Sun, Y.; Wu, W.; Loy, C. C.; Wang, X.; and Liu, Z. 2021. Pose-controllable talking face generation by implicitly modularized audio-visual representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4176–4186.

Zhou, Y.; Han, X.; Shechtman, E.; Echevarria, J.; Kalogerakis, E.; and Li, D. 2020. Makelttalk: speaker-aware talking-head animation. *ACM Transactions On Graphics (TOG)*, 39(6): 1–15.

Zhu, H.; Wu, W.; Zhu, W.; Jiang, L.; Tang, S.; Zhang, L.; Liu, Z.; and Loy, C. C. 2022. CelebV-HQ: A Large-Scale Video Facial Attributes Dataset. In *ECCV*.