2024 18th International Conference on Automatic Face and Gesture Recognition (FG)

EAT-Face: Emotion-Controllable Audio-Driven Talking Face Generation via Diffusion Model

Haodi Wang[1], Xiaojun Jia[2] and Xiaochun Cao[1†]
[1] Sun Yat-sen University, China [2] Nanyang Technological University, Singapore

Fig. 1. **Emotion-controllable audio-driven talking face generation.** Our framework enables the synthesis of Emotion-controllable Audio-driven Talking Faces (**EAT-Face**). By using learned joint emotion-visual embedding, the EAT-Face can manipulate the facial emotion of the generated talking face based on the audio and emotion-text input.

***Abstract*— Audio-driven talking face generation is a promising task with a lot of attention. Despite abundant efforts are devoted to video quality and lip synchronization, most existing works do not take the unignorable aspect of facial emotional expression into account during generation. In this paper, we propose an Emotion-controllable Audio-driven Talking Face generation framework called *EAT-Face* that enables us to control multiple types of emotions. Specifically, the proposed method consists of a *Talking Face Reconstructor* (TFR) and a *Facial Emotion Controller* (FEC), utilizing fused multimodal information including audio signals, visual images, and textual emotions for synthesis. Firstly, TFR predicts face images synchronized with given audios from random noises, leveraging external guidances comprised of audio features, character references, and face masks as conditions. Then, FEC further manipulates the facial emotions based on TFR, leveraging the emotion embeddings extracted from emotion texts. However, a semantic misalignment problem lies in the emotion-texts and character images. To tackle this issue, we additionally propose a strategy called *joint Emotion-Visual Embedding* (EVE) to mitigate the misalignment. In this way, the proposed EAT-Face is captive to control emotion more precisely. Extensive experiments involving both objective evaluations and subjective investigations demonstrate the effectiveness of our framework in synthesizing high-fidelity and emotional talking face videos.**

I. Introduction

Artificial intelligence has achieved excellent performance in many fields [27][71]. Specially, audio-driven talking face generation is a challenging but captivating task in the field of artificial intelligence generation [14], brimming with potential and promising prospects. Recently, it has been preliminarily applied in many domains such as virtual anchor [28][42][61], character re-dubbing [30][67], video conferencing [66], animation film [73] and so on.

†Corresponding author.

Previous works have contributed efforts towards the development of high-fidelity talking face generation, predominantly relying on Generative Adversarial Networks (GAN) [11][12]. However, there are certain limitations in the existing methods. For instance, it is annoying for researchers to face with instability of GAN's training. In the flourishing period of popular artificial intelligence-generated content (AIGC), the emergence of Diffusion Models (DM) provides greater freedom and possibilities for AI creation. In comparison with GAN, the advantages of DM are embodied in its superior training stability, as well as its outstanding visual effect in many tasks. Recently, several works [2][59][53] have employed DM for audio-driven talking face generation.

It is worth noting that characters usually talk with certain emotions in the real world. However, most existing methods pay more attention to the generation quality and audio-visual synchronization, while ignoring facial emotional expressions. Only a few works with GAN attempt to consider this aspect so far, and the works with DM still overlook the incorporation of emotions during generation.

The question that stimulates our thinking is how to achieve emotion control in the generation of DM. Analogous to the text-to-image tasks, an intuitive solution is to utilize texts containing emotional words as guiding conditions for generation. However, we observe that this idea does not meet the expectations in terms of the strength and effect of emotion control. We speculate that this issue may arise from a misalignment problem between emotion-text and image content. In other words, the entire image content is described by the text prompt used in DM, while emotion is solely expressed on the facial region, leading to an insufficient correlation between them.

979-8-3503-9494-8/24/$31.00 ©2024 IEEE

In this paper, we propose a crafted framework called **EAT-Face** for the generation of talking faces, aiming to address the aforementioned challenges. The framework comprises two pivotal components: **Talking Face Reconstructor (TFR)** and **Facial Emotion Controller (FEC)**. Specifically, within the **TFR**, we leverage the latent conditional diffusion model as the foundation for the generation of general faces, i.e. faces without emotion. Meanwhile, the usage of this module avoids training complexities associated with GAN. As guidance conditions of DM, the driven audio, identity reference, and mask are introduced to facilitate audio-visual synchronization, identity preservation, and high-quality generation. As for the **FEC**, we devise a shared embedding space that enables us to obtain learned joint embedding of emotion-text and visual-image through contrastive learning, which better alleviates the problem of misalignment. The acquired joint embeddings are then utilized as emotion conditions to TFR through an emotion ControlNet for precise facial emotion manipulation. With these designs, we enhance the controllability over expressed emotions upon generated faces, while maintaining the desired high visual quality.

Sufficient experiments demonstrate the controllability of our method for talking face generation. As shown in Fig. 1, the proposed method effectively generates natural talking videos with different emotions. Our main contributions are summarized as follows:

- We propose an Emotion-controllable Audio-driven Talking Face generation framework based on the diffusion model, called EAT-Face, for synthesizing high-fidelity, audio-visual synchronized talking face videos with photo-realistic expressions.
- We design a shared representation space for learning joint embedding of emotion-text and visual content, better improving the alignment between them and leading to precise emotion control.
- We propose a facial emotion control module to provide more effective emotional semantic conditions for the diffusion model.
- Extensive experiment results demonstrate that our proposed method can not only achieve satisfactory and comparable visual quality to other methods, but also provide more delicate emotional manipulation of facial details that are lacking in those methods.

II. Related Work

A. Diffusion Models for Visual Generation

In the field of visual generation, the Denoising Diffusion Probabilistic Model (DDPM) [17] stands as the pioneering diffusion model, with most subsequent works [18][41][46][56][58] building upon its foundation. Various samplers [29][32][56][57][58] are proposed to accelerate the generation of the diffusion model, among which the Denoising Diffusion Implicit Model (DDIM) [56] is a typical method using a deterministic denoising process with fewer sampling steps.

The Latent Diffusion Model (LDM) [46] transfers the process of diffusion and denoising from pixel space to latent space, reducing computational consumption while keeping the quality of the generated samples. Following this, [9] demonstrates that diffusion models exhibit significant potential for beating GAN, which has demonstrated remarkable achievements across various domains afterward, such as image super-resolution [18][51], image inpainting [35][49], text-to-image synthesis [40][45][50], text-to-video synthesis [16][43][55], 3D point cloud [7][36], and so on.

More recently, re-learning approaches for large models that focus on different specific tasks alleviate the challenges associated with training diffusion models. For instance, Low-Rank Adaptation of Large Language Model (LoRA) [20] is proposed to learn a specific image style based on existing diffusion models; ControlNet [70] introduces a trainable branch copied from diffusion models to support additional semantic mappings and locks the original weights from pretrained diffusion models, enabling more delicate control over synthesized images. Similar works such as [39][48] further enhance the generation capability of diffusion models.

Inheriting the advantages of LDM and its re-learning ability, in this paper, we employ the LDM for talking face generation to achieve high-fidelity synthesis results.

B. Talking Face Generation

Existing works on audio-driven talking face generation can be roughly divided into three categories according to the basic backbone they employ, namely based on Generative Adversarial Networks (GAN), Neural Radiance Fields (NeRF), and Diffusion Models (DM) respectively.

1) GAN-based Methods: It is prevalent for this task to employ GAN [11][12][21][23][22], commonly with the assistance of intermediate representations such as facial landmarks and 3D facial coefficients. Among the methods based on landmarks such as [5][26][33][72][60][73], several works [26][60][33] utilize Recurrent Neural Networks (RNN) [37] or Long Short Term Memory networks (LSTM) [19] to learn the mapping from audio to landmark movements. In the methods of 3D facial coefficients, the 3D Morphable Model (3DMM) [3] is utilized in [31][68][61][69] to predict facial parameters for synthesizing talking faces.

2) NeRF-based Methods: NeRF [38] provides another solution for talking face generation, which is used in AD-NeRF [15] and DFRF [52] to get better results.

3) DM-based Methods: Most recently, the diffusion model [17] has provided a brand-new framework for talking face generation. [59] defines the synthesis as an autoregressive task in pixel space, while its generation speed is constrained. DiffTalk [53] and DAETalker [10] are trained in a latent space using conditional latent diffusion models. In [53], DDIM [56] is employed to generate video frames sequentially during sampling, while [10] proposes an innovative parallel strategy based on DDIM to speed up sampling.

As for emotion control, only a few GAN-based methods [13][25][26][64] consider this issue. For instance, [25] proposes an emotion-aware motion model to synthesize emotional faces, [26] proposes an approach to decouple

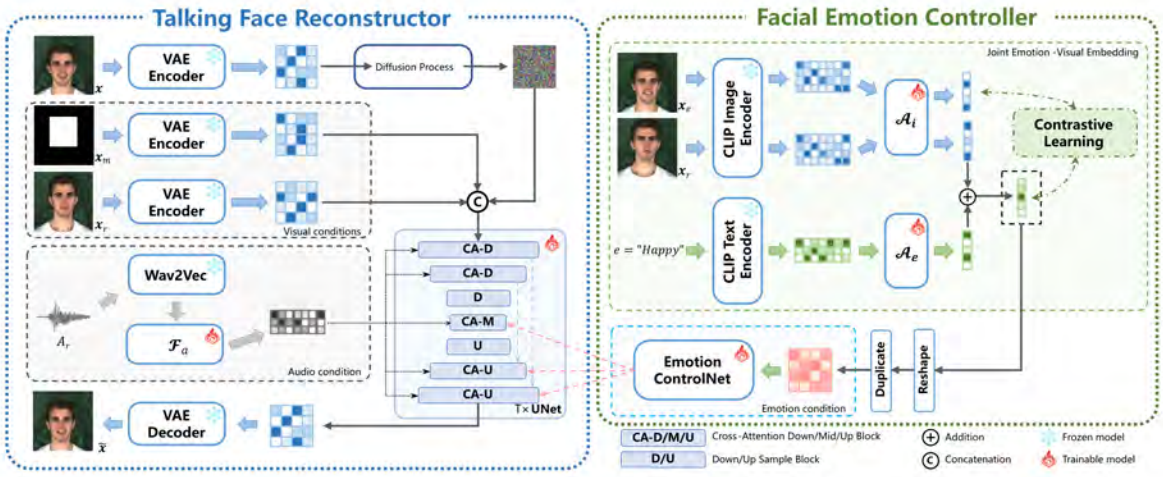

Fig. 2. **Overview of our EAT-Face framework.** It consists of two parts: (1) **Talking Face Reconstructor (TFR).** The latent diffusion-based reconstructor takes audio, mask, and identity conditions as guidance to generate high-fidelity talking face frames with general emotion. (2) **Facial Emotion Controller (FEC).** The ControNet-based controller takes emotional conditions as guidance to further control facial emotion based on face images synthesized from TFR. Meanwhile, the module of joint emotion-visual embedding is used to address the misalignment between face image and emotion.

emotional features from audio signals, and [13] focuses on realistic-looking emotional talking videos generation.

However, no research has yet mentioned how to control character emotion in DM-based models. In this work, we propose for the first time to control the emotion in the diffusion model by using joint emotion-visual embedding as the guidance condition.

III. Methodology

A. Overview

Given a reference image $\boldsymbol{x}_r$ of a specific character, a sequence of audio A_r, and an emotion-text e, we aim to generate video frames $\tilde{\boldsymbol{x}} = \{\tilde{\boldsymbol{x}}_1, ..., \tilde{\boldsymbol{x}}_N\}$, where $\tilde{\boldsymbol{x}}_i$ is the i-th predicted frame that is lip-synchronized with the audio and corresponding with the emotion while keeping the original identity. An overview of our proposed EAT-Face is depicted in Fig. 2. Two main modules named Talking Face Reconstructor (TFR) and Facial Emotion Controller (FEC) are included in the framework, which are both utilized simultaneously to generate images at inference time.

In the coming sections, in section III-B we begin by briefly reviewing the diffusion model, which is the basis of our EAT-Face. Then, we respectively detail the architecture of TFR and FEC in sections III-C and III-D. Lastly, we introduce the adopted inference strategy in III-E.

B. Preliminary: Diffusion Model

The diffusion model is utilized as the foundation of our EAT-Face. The pure diffusion model takes the noisy image $\tilde{\boldsymbol{x}}_t$ from the previous timestep and the time embedding t as inputs at each step, predicting the residual noise and then obtaining the denoised image $\tilde{\boldsymbol{x}}_{t-1}$. The process can be formulated as:

$$\tilde{\boldsymbol{x}}_{t-1} = \tilde{\boldsymbol{x}}_t - \epsilon_\theta(\tilde{\boldsymbol{x}}_t, t), \quad t \in \{T, ..., 2, 1\}, \tag{1}$$

$$\tilde{\boldsymbol{x}}_T \sim (0, \boldsymbol{I}), \tag{2}$$

where ϵ_θ is a trainable denoising UNet [47] model.

The text c_{text} is usually employed as an additional input to guide denoising to exert further control over the generated content or style. The optimization objective of the diffusion model can thus be mathematically formulated as follows:

$$\mathcal{L}^{simple} = \mathbb{E}_{\boldsymbol{x},\boldsymbol{\epsilon}\sim\mathcal{N}(0,\boldsymbol{I}),t,c_{text}}[\|\boldsymbol{\epsilon} - \epsilon_\theta(\tilde{\boldsymbol{x}}_t, t, c_{text})\|_2^2], \tag{3}$$

C. Talking Face Reconstructor

In this section, we introduce the Talking Face Reconstructor (TFR) based on the conditional diffusion model, aiming to generate audio-synchronized and identity-preserving face images. Notably, emotion control is not considered within this part.

Both the diffusion and denoising processes are conducted in the latent space to enhance training efficiency and reduce computational overhead, following [46]. We employ the pretrained encoder $\mathcal{E}$ and decoder $\mathcal{D}$ from the Variational Auto-Encoder (VAE) [62] as a bridge connecting latent space and pixel space. For a face image $\boldsymbol{x} \in \mathbb{R}^{3\times H\times W}$ extracted from a video, where H and W denote the height and width of the original image respectively, we utilize $\mathcal{E}$ to encode it into corresponding latent image $\boldsymbol{z} = \mathcal{E}(\boldsymbol{x}) \in \mathbb{R}^{4\times h\times w}$, where h and w are determined by a down-sampling f of VAE i.e. $f = H/h = W/w$. The diffused $\boldsymbol{z}_T \in \mathbb{R}^{4\times h\times w}$ obtained by (1) is then denoised by UNet to yield the predicted latent original image $\tilde{\boldsymbol{z}}_0 \in \mathbb{R}^{4\times h\times w}$. We finally obtain the desired generated image $\tilde{\boldsymbol{x}}$ in pixel space after decoding, i.e. $\tilde{\boldsymbol{x}} = \mathcal{D}(\tilde{\boldsymbol{z}}_0) \in \mathbb{R}^{3\times H\times W}$.

Here, three extra conditions as elaborated below are considered to guide the generated face.

1) Audio Condition: The audio condition is introduced as guidance to drive facial movements, such as lip shape, blinking, muscle orientation, etc. For a raw audio A_r, we firstly employ the pretrained Wav2Vec [1] model $\mathcal{W}$ to extract its phoneme feature $\boldsymbol{A} = \mathcal{W}(A_r) = [\boldsymbol{a}_0; \boldsymbol{a}_1; ...; \boldsymbol{a}_{S-1}] \in$

$\mathbb{R}^{S\times D}$, where S is the audio frame length and D is the feature dimension. To be more concerned about the correspondence between mouth movements and phonemes rather than the audio characteristics associated with specific characters, we refrain from utilizing conventional audio features such as Mel-spectrum. It should be noted that the length of audio frames differs from that of video frames. To maintain consistency in audio features across a sequence of frames, a sliding window of size 3 is employed to determine the audio features for each image frame. Specifically, the audio features for the i-th frame ($i = 1, 2, ..., N$) consist of $[\boldsymbol{a}_{u-1}; \boldsymbol{a}_u; \boldsymbol{a}_{u+1}]$, where $u = \lfloor S(i-1)/N \rfloor$, and the vacant portion is padded with 0 when $u = 0$ or $u = S-1$. We subsequently utilize a $\mathcal{F}_a$ that is comprised of 1D-convolution layers and multi-layer perceptron (MLP) to acquire audio condition embeddings $\boldsymbol{c}_a$ from the features. Replacing the location of prompt text in the original diffusion model, these embeddings inject audio information into the cross-attention layer of UNet serving as keys and values.

2) Mask Condition: According to [69], speech audios exhibit not only a strong direct correlation with lip movements but also an indirect correlation with other facial movements, so the modification of the whole face is needed. Additionally, we observe that the facial area shows more variation than other areas like the background when a character is speaking, implying that it is unnecessary to manipulate all regions of the image. Inspired by [35], we introduce a mask condition to guide the focus area generated. Our approach differs from previous methods in that we use a mask that covers most of the face region, instead of just the mouth area [2] or the lower half of the image [53]. Specifically, we leverage MediaPipe [34] for facial landmarks detection and select specific points #127, #151, #152, and #356 to determine the boundary of the mask, resulting in obtaining a mask image $\boldsymbol{x}_m$. The mask condition embedding $\boldsymbol{z}_m$ is calculated via the VAE encoder, i.e. $\boldsymbol{z}_m = \mathcal{E}(\boldsymbol{x}_m) \in \mathbb{R}^{4\times h\times w}$.

3) Identity Condition: The identity information of the face image is almost entirely corrupted after the diffusion process. To ensure consistency between the generated and original faces, we randomly select a frame $\boldsymbol{x}_r$ from the video as the reference image to guide denoising. We then acquire the identity reference condition $\boldsymbol{z}_r = \mathcal{E}(\boldsymbol{x}_r) \in \mathbb{R}^{4\times h\times w}$.

Considering that $\boldsymbol{z}_t$, $\boldsymbol{z}_m$, and $\boldsymbol{z}_r$ all represent implicit spatial images, we adopt the practice in [53] to concatenate them along the channel as the input for the UNet model, which can be formulated as follows:

$$\mathbf{x}_{in} = \boldsymbol{z}_t \oplus_c \boldsymbol{z}_m \oplus_c \boldsymbol{z}_r, \tag{4}$$

where $\oplus$ denoted the operation of concatenation. The $\boldsymbol{z}_m$ and $\boldsymbol{z}_r$ are consolidated as a visual condition $\boldsymbol{c}_v = \{\boldsymbol{z}_m, \boldsymbol{z}_r\}$.

Based on the aforementioned additional guidance conditions, the training objective for TFG is defined as follows:

$$\mathcal{L}^{TFG} = \mathbb{E}_{\boldsymbol{z},\boldsymbol{\epsilon}\sim\mathcal{N}(0,\boldsymbol{I}),t,\boldsymbol{c}_a,\boldsymbol{c}_v}[\|\boldsymbol{\epsilon} - \epsilon_\theta(\tilde{\boldsymbol{z}}_t, t, \boldsymbol{c}_a, \boldsymbol{c}_v)\|_2^2]. \tag{5}$$

D. Facial Emotion Controller

1) Joint Emotion-Visual Embedding: Inspired from the contrastive learning in CLIP [44], we design a shared embedding space to acquire the joint emotional representation of textual emotion and visual image, and we call it joint Emotion-Visual Embedding (EVE). We train two adapter networks $\mathcal{A}_v$ and $\mathcal{A}_e$ to map visual images and emotional texts into an embedding space respectively. Both of them are composed of multiple alternating convolution layers, max-pooling layers, and activation layers, and then end with an MLP.

Firstly, we leverage the CLIP image encoder $\mathcal{E}_{CI}$ to encode the facial image $\boldsymbol{x}_e$ with emotional expression and pass it to $\mathcal{A}_v$, obtaining the visual representation $\boldsymbol{h}^v \in \mathbb{R}^{hw}$. Then, the corresponding emotional text e is encoded by the CLIP text encoder $\mathcal{E}_{CT}$ and we obtain the text representation $\boldsymbol{h}^e \in \mathbb{R}^{hw}$ by $\mathcal{A}_e$. To preserve person-specific identity information in the text representation, a corresponding emotionless facial image $\boldsymbol{x}_r$ is encoded by $\mathcal{A}_v$ and fused into $\boldsymbol{h}^e$ through weighted addition as an identity-based feature. This process can be formulated as follows:

$$\boldsymbol{h}^v = \mathcal{A}_v(\mathcal{E}_{CI}(\boldsymbol{x}_e)), \tag{6}$$

$$\boldsymbol{h}^e = \alpha\mathcal{A}_e(\mathcal{E}_{CT}(e)) + (1-\alpha)\mathcal{A}_v(\mathcal{E}_{CI}(\boldsymbol{x}_r)), \tag{7}$$

where α denotes the emotion weight. During training, a minibatch consists of N emotion-visual representation pairs $(\boldsymbol{h}_i^e, \boldsymbol{h}_i^v)$ and does not contain the same emotion. We aim to maximize the similarity between emotion-matching pairs and minimize the similarity between mismatched pairs so that the emotion-text can be aligned with the corresponding face image as much as possible. The optimization objective consists of emotion-to-visual contrastive loss and visual-to-emotion contrastive loss. For the i-th pair, the emotion-to-visual contrastive loss is formulated as:

$$\mathcal{L}_i^{e\to v} = -\log\frac{\exp(\langle \boldsymbol{h}_i^e, \boldsymbol{h}_i^v\rangle/\tau)}{\sum_{k=1}^N \exp(\langle \boldsymbol{h}_i^e, \boldsymbol{h}_k^v\rangle/\tau)}, \tag{8}$$

and the visual-to-emotion contrastive loss is formulated as:

$$\mathcal{L}_i^{v\to e} = -\log\frac{\exp(\langle \boldsymbol{h}_i^v, \boldsymbol{h}_i^e\rangle/\tau)}{\sum_{k=1}^N \exp(\langle \boldsymbol{h}_i^v, \boldsymbol{h}_k^e\rangle/\tau)}, \tag{9}$$

where $\langle \boldsymbol{a}, \boldsymbol{b}\rangle$ denotes the cosine similarity between representation vector $\boldsymbol{a}$ and $\boldsymbol{b}$, and τ denotes the temperature coefficient, which is set to 0.2 in our experiments.

For a minibatch, the total loss function for EVE is formulated as:

$$\mathcal{L}^{e\leftrightarrow v} = \frac{1}{2N}\sum_{i=1}^N(\mathcal{L}_i^{e\to v} + \mathcal{L}_i^{v\to e}). \tag{10}$$

2) Emotion Controlling via ControlNet: For the manipulation of facial emotion, an intuitive approach is to fuse emotion features directly with audio features to guide denoising. However, we find that such a direct way does not work as expected in a preliminary experiment. The ControlNet [70] further enhances the control over the pretrained diffusion model by various forms of conditions while preserving the generative capability of the existing diffusion model. Inspired by it, we particularly adopt a ControlNet, known as Emotion ControlNet, to achieve emotion control for facial images

TABLE I
QUANTITATIVE COMPARISON WITH OTHER METHODS

Dataset	Method	w/Emo.	Reconstruction Quality			Perceptual Similarity		Synchronization		ID Preservation
			PSNR↑	SSIM↑	CPBD↑	LPIPS↓	FID↓	LSE-D↓	LSE-C↑	CSIM↑
MEAD	MakeItTalk[73]		25.7035	0.7914	0.1215	0.1531	50.9649	9.2916	5.3394	0.7648
	StyleHEAT[68]		17.9452	0.5063	0.3354	0.2321	195.2083	13.9266	2.3095	0.6000
	IP-LAP[72]		32.1435	**0.9448**	0.2538	0.0458	28.6454	**8.8531**	**6.2697**	0.8386
	EAMM[25]	✓	18.1904	0.6312	0.1266	0.2317	122.5730	12.3489	2.7372	0.6010
	EmoGen[13]	✓	16.2721	0.6028	0.3027	0.2304	56.3227	12.1095	2.6248	0.5848
	EAT-Face(Ours)	✓	**32.6131**	0.9275	**0.3803**	**0.0274**	**15.7736**	9.1911	5.5614	**0.8460**
CREMA-D	MakeItTalk[73]		23.5662	0.7740	0.1143	0.0756	19.8448	9.4263	3.2461	0.7287
	StyleHEAT[68]		19.2144	0.6658	0.0892	0.1309	63.9966	8.8267	3.4079	0.5152
	IP-LAP[72]		32.4785	**0.9476**	0.4725	0.0167	18.5343	7.2367	4.2880	0.7208
	EAMM[25]	✓	19.3724	0.6829	0.2579	0.1487	197.2419	8.3457	3.9349	0.6179
	EmoGen[13]	✓	22.4822	0.7412	0.5154	0.0803	22.5024	7.4055	4.4748	0.7402
	EAT-Face(Ours)	✓	**34.3401**	0.9325	**0.5180**	**0.0147**	**9.9181**	**6.7923**	**4.6093**	**0.7562**

especially. Different from conditions such as skeleton image and depth image used in common ControlNet, the Emotion ControlNet takes EVE as extra semantic control conditions.

Specifically, for the learned EVE which fuses emotion feature with character visual information, we reshape it into a 2D feature map and replicate it 4 times along the channel as the emotion condition $\boldsymbol{c}_e \in \mathbb{R}^{4\times h\times w}$. This operation is performed to suit the dimensionality of $\boldsymbol{c}_e$ with that of the latent space and then correctly convey it to ControlNet as condition input. The output of our Emotion ControlNet is subsequently fed into the cross-attention layers of intermediate blocks and up-sampling blocks within UNet. During training, we first train Emotion ControlNet separately while keeping the weights of other networks frozen. Once the training is complete, we unfreeze the weight of the UNet part and fine-tune it through Emotion ControlNet.

To summarize, with the design of our emotion controller, we fine-tune the diffusion model trained in the previous section, and the optimization objective is adapted as follows:

$$\mathcal{L} = \mathbb{E}_{\boldsymbol{z},\boldsymbol{\epsilon}\sim\mathcal{N}(0,\boldsymbol{I}),t,\boldsymbol{c}_a,\boldsymbol{c}_v,\varphi_\theta}[\|\boldsymbol{\epsilon} - \epsilon_\theta(\tilde{\boldsymbol{z}}_t, t, \boldsymbol{c}_a, \boldsymbol{c}_v, \varphi_\theta(\boldsymbol{c}_e))\|_2^2], \tag{11}$$

where φ_θ denotes the trainable Emotion ControlNet.

E. Sampling

In the inference phase, our EAT-Face receives a reference image $\boldsymbol{x}_r$, an audio clip A_r, and an emotion-text e as inputs, and then outputs a sequence of image frames. Note that the mask condition $\boldsymbol{x}_m$ is obtained automatically, and the $\boldsymbol{x}_e$ used in the training phase is no longer used during inference. The text undergoes sequential processing through the CLIP text encoder, trained adapter $\mathcal{A}_e$, and Emotion ControlNet to obtain emotional condition, which is utilized for guiding UNet denoising. We conduct generation with DDIM [56], which yields high-quality samples with fewer iterations compared to DDPM [17]. Following the practice of [10], we adopt parallel sampling instead of progressive sampling to accelerate the generation process. Specifically, all the image frames in a certain video are initialized with a shared state, i.e. the same noise $\boldsymbol{z}_T \sim \mathcal{N}(0, \boldsymbol{I})$, resulting in continuous changes when conduct sampling in parallel. Additionally, the frame interpolation technology [24] is utilized for inter-frame smoothing in the post-processing stage to effectively address potential issues such as frame skipping and jitter.

IV. EXPERIMENTS

A. Experimental Setup

1) Dataset: In our experiments, we utilize two emotional audio-visual datasets: **MEAD** [65] with 8 emotions and **CREMA-D** [4] with 6 emotions. For MEAD, videos with medium-intensity emotion and a front-camera view are specifically selected for our experiments. We randomly sample 1,472 videos from this subset for training and testing, resulting in approximately 93k and 94k human images in the training and test datasets, respectively. For CREMA-D, we randomly sample 1,820 videos, resulting in approximately 68k and 62k human images in the training and testing datasets.

2) Data Preprocess: The original videos are cropped into square size with the character's face centered. Subsequently, image frames are extracted at 25fps and saved at 256×256 resolution. The audios are extracted from corresponding videos and resampled at 16,000Hz. As for emotion texts, we transform the emotion labels associated with the videos into prompt-style forms like $* \; emotion$, where $* \in \{angry, contempt, disgusted, fear, happy, neutral, sad, surprised\}$. Note that CREMA-D [4] does not contain $contempt$ or $surprised$. We adopt the standardized format for emotion labels as emotion control text.

3) Implementation Details: The UNet is trained based on a 2D conditional UNet architecture with the size of latent space setting to 32. During training, the batch sizes and learning rates for UNet, Emotion ControlNet, and EVE are respectively set to 48, 12, 8, and 1e-5, 1e-5, 1e-3. The optimizer is AdamW with a weight decay of 1e-3 and warm steps of 500. The UNet is trained on an 80G Nvidia A100 GPU and the other modules are on a 24G Nvidia GeForce

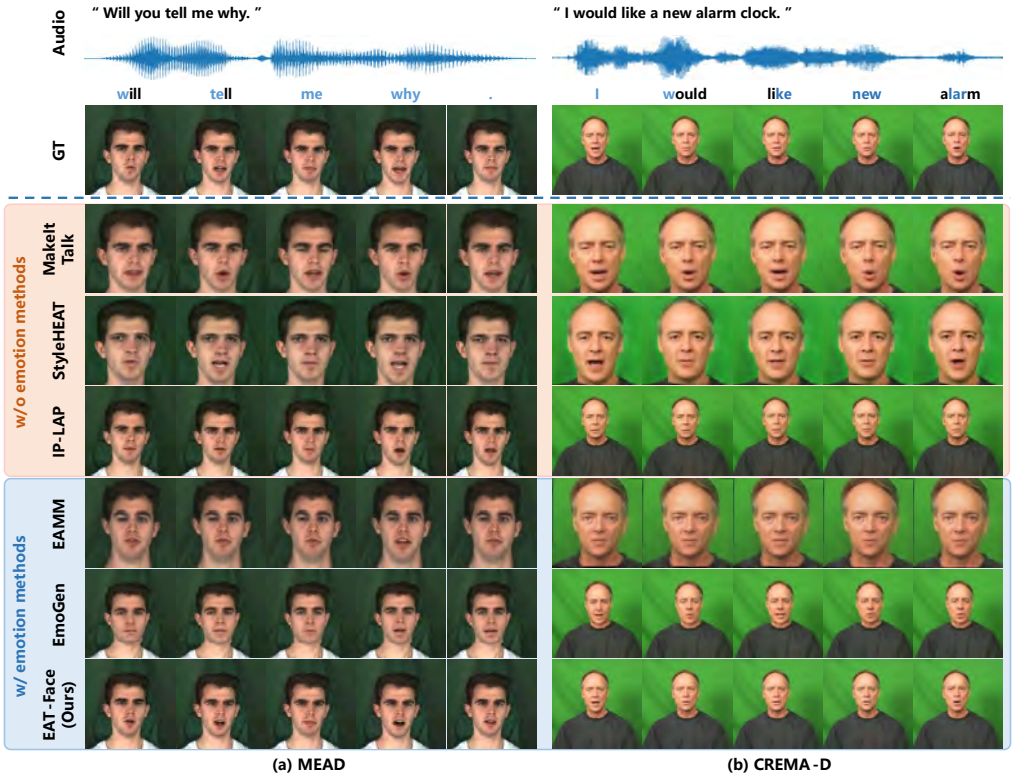

Fig. 3. **Qualitative comparison with other methods.** Visually our method produces better results than other methods MakeItTalk [73], StyleHEAT [68], IP-LAP [72], EAMM[25], and EmoGen[13] on both MEAD [65] and CREMA-D [4].

RTX 3090 GPU. During inference, DDIM with 100 denoise steps is used as the sampler.

B. Methods Comparison

In this section, we conduct comparative experiments quantitatively and qualitatively between our EAT-Face and several baseline methods [13][25][68][72][73]. **MakeItTalk** [73] is a classical method for talking face generation according to 3D landmarks. **StyleHEAT** [68] leverages 3D facial coefficients to realize one-shot synthesis. **IP-LAP** [72] is a more recent method that utilizes 2D landmarks to animate a person based on given audio. **EAMM** [25] and **EmoGen** [13] can further generate talking portraits with desired emotion types. We utilize their released checkpoints for evaluation in our experiments.

1) Quantitative Comparison: To assess the image reconstruction quality, we employ three metrics, Peak Signal-to-Noise Ratio (**PSNR**), Structural Similarity (**SSIM**), and Cumulative Probability of Blur Detection (**CPBD**). Additionally, to measure perceptual similarity, we utilize two metrics Learned Perceptual Image Patch Similarity (**LPIPS**) and Fréchet Inception Distance (**FID**) that align better with visual characteristics. Furthermore, lip-audio synchronization is evaluated using Lip Sync Error-Distance (**LSE-D**) and Lip Sync Error-Confidence (**LSE-C**) from SyncNet [6], while identity similarity of the character is measured through the Cosine Similarity score (**CSIM**) from ArcFace [8].

The quantitative results are presented in Table I. As can be seen, the proposed EAT-Face shows excellent quality in image reconstruction, particularly on the metrics of PSNR and CPBD. In terms of perceptual similarity, our method is superior to other methods. Particularly on the FID, our method significantly surpasses MakeItTalk, StyleHEAT, IP-LAP, EAMM, and EmoGen by an average of 59.54%, 88.21%, 45.71%, 91.14%, and 63.96% respectively, indicating a stronger correlation with high-quality images. It can be attributed to utilizing the diffusion model as the foundation for generation which exhibits superior capability compared to GAN resulting in higher quality. Regarding lip-audio synchronization, on the MEAD dataset, our method is marginally inferior to IP-LAP, and on par with MakeItTalk; on the CREMA-D dataset, our method gets the best score, demonstrating that EAT-Face possesses the ability to maintain synchronization. Moreover, our method achieves higher CSIM, which means EAT-Face possesses a better ability to preserve identity information. The reason is that our method makes full use of identity encoding as one of the conditions to guide generation.

2) Qualitative Comparison: Fig. 3 displays two examples of different methods. From the perspective of visual perception, the results of EAT-Face are closer to the ground truth. Firstly, similar to MakeItTalk and IP-LAP, our method preserves more original identity information of the figure to a greater extent than the rest 3 methods which have a larger deviation from the ground truth. Besides, StyleHEAT's smoother face reduces its authenticity, and EAMM is easy to lead to a phenomenon of asymmetrical eyes. Secondly,

Fig. 4. **Qualitative evaluation results of emotion control on MEAD.** Different columns denote different emotions. The 1st row is the ground truth images, and the 2nd row is the face images generated by our EAT-Face.

TABLE II
QUANTITATIVE EVALUATION RESULTS OF EMOTION CONTROL

Emotion Type	Emo-Acc.	Emotion Type	Emo-Acc.
Angry	82.51%	Happy	81.41%
Contempt	80.02%	Neutral	78.66%
Disgusted	75.19%	Sad	81.99%
Fear	65.65%	Surprised	51.13%
Average Emo-Acc.		**74.57%**	

apparent and arrestive artifacts around the person's contour exist in StyleHEAT and EmoGen. Our method successfully avoids generating these fuzzy or blunt details. In terms of mouth movement, the separation between the upper and lower lips of the talker may occasionally be not obvious in IP-LAP, leading to an unnatural appearance. Our results reduce this issue. Additionally, Fig. 3 shows the mouth shape corresponding to some words, and more accurate visual results appear in our EAT-Face. For example, the corresponding mouth shape for the phonetic sign of /w/ follows the usual situation, and when the sentence ends, the character's lip is closed. Finally, compared to other methods, our results better preserve teeth shown during talking, resulting in more authentic generated images.

C. Emotion Control Evaluation

To verify the effectiveness of EAT-Face on emotional face generation, we conducted quantitative and qualitative validation on the MEAD dataset.

1) Quantitative Results: To evaluate the control capability on facial emotions of EAT-Face quantitatively, we additionally train an emotion classifier based on VGG-16 [54] for emotion classification of generated images. The accuracy of predicted results by the classifier, marked as **Emo-Acc.**, serves as the metric for evaluating our method's ability to control emotions, where a higher Emo-Acc. indicates superior emotional control.

The results are presented in Table II. It can be seen that the classifier accurately identifies the emotion of over 74.6% of image frames on average in the generated videos, affirming the effectiveness of our method for emotion control. Notably, compared to emotions like fear and surprise, our method demonstrates more proficiency in other emotions such as happiness, sadness, anger, and contempt.

2) Qualitative Results: The visual results of EAT-Face under different emotion-text settings are presented in Fig. 4. The generated faces exhibit expressions that are close to the semantic emotion of the specified text while preserving identity information from reference images, indicating the visual effectiveness of our method in emotion control.

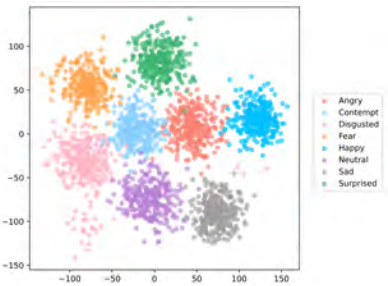

Fig. 5. **The t-SNE** [63] **visualization of emotion-control condition embeddings.** Different colors indicate different emotion types.

Additionally, we utilize t-SNE [63] to visualize the emotion conditions in Fig. 5. The clear distinction between different emotion clusters demonstrates the effectiveness of our emotion control.

D. User Study

We invite 38 participants to conduct the user study, and they are divided into 2 groups for different investigations.

1) User study on EAT-Face: The first group is asked to subjectively evaluate the generated videos by our EAT-Face from 4 aspects: the tendency of emotions(**Tend.**), the degree of emotional expression (**Emo-De.**), the reality of videos (**Real.**), and the naturalness of characters(**Nat.**) respectively. The metric of tendency is calculated by participants' voting, and the rest metrics are scored by participants from the range 1 to 10, with higher scores indicating better quality.

The results are depicted in Fig. 6. As can be seen, the average percent of the emotional tendency of participants reaches 73.2%, demonstrating a discernible emotional inclination among the characters featured in the generated videos. The orientations for four emotion categories, namely contempt, happy, neutral, and sad, are relatively high with an average of approximately 80%, indicating that EAT-Face exhibits strong reconstruction capability for these kinds of emotions. The average degree of emotional expression in the generated video stands at 7.192, which aligns well with our expectations due to the medium-intensity emotion selected in the training set. Subjectively speaking, both the reality of videos and the naturalness of characters fall within normal levels but are influenced by various factors such as frame continuity and facial dynamic changes.

2) User Study among Methods: The second group is asked to subjectively score the videos generated by different methods from 4 aspects: the expression of emotion(**Emo.**), the visual quality(**Vis-Q.**), the lip-audio synchronization(**Sync.**), and the identity preservation(**ID-P.**). The score ranges from 1 to 10, in which higher means better. The investigation results shown in Table III demonstrate that our EAT-Face is better than other methods in all 4 aspects mentioned above, especially with the significant improvement of emotional expression.

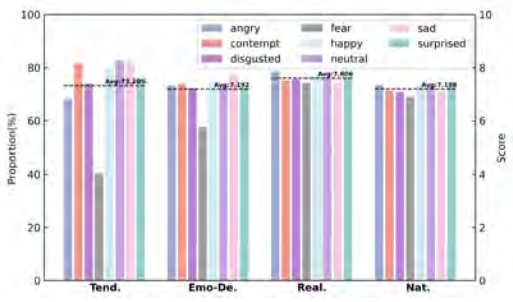

Fig. 6. **Statistical results of the user study on EAT-Face.** Different colors denote different emotions, and the black dashes denote the average statistical results of corresponding metrics.

TABLE III

USER STUDY AMONG DIFFERENT METHODS

Method	Emo.↑	Vis-Q.↑	Sync.↑	ID-P↑
MakeItTalk[73]	5.730	6.395	7.020	8.395
StyleHEAT[68]	5.813	6.148	5.605	7.480
IP-LAP[72]	5.648	7.458	7.543	8.438
EAMM[25]	6.357	5.429	7.286	7.143
EmoGen[13]	6.286	5.500	7.714	7.500
EAT-Face(Ours)	**8.190**	**7.543**	**7.918**	**8.730**

TABLE IV

ABLATION STUDY ON EMOTION-RELATED MODULES

Method	Components				Metrics	
	DM	EVE	FEC	Emo	Emo-Acc.↑	Tend.↑
w/o EVE	✓	✗	✓	✓	62.88%	53.76%
w/o FEC	✓	✗	✗	✓	15.03%	19.24%
w/o Emo	✓	✗	✗	✗	13.51%	4.83%
Ours	✓	✓	✓	✓	**74.57%**	**73.20%**

E. Ablation Study

We conduct ablation studies on the MEAD dataset.

1) Impact of joint EVE: As shown in Fig. 7, we calculate the cosine similarities between visual image embeddings and emotion-text embeddings before and after joint EVE. It can be seen that the joint EVE module significantly reduces the distance between these two embeddings in the representation space. This indicates that the proposed joint EVE does contribute to aligning visual and emotional information.

2) Effectiveness of Emotion Control Modules: We attempt to explore the effects of the joint EVE and the FEC module. We try to remove or modify certain components as the variation of our proposed method for the ablation study. Three variants are employed: (*i*) the EVE is removed and the encoded emotion-text is fed to ControlNet directly (mark as w/o EVE); (*ii*) the FEC is removed and the encoded emotion-text is concatenated with audio signal as the hidden state condition (mark as w/o FEC); (*iii*) no emotional condition is used (mark as w/o Emo). As can be seen from the results in Table IV, the removal of any part will lead to a decline in the generation effect, indicating the effectiveness of the proposed modules.

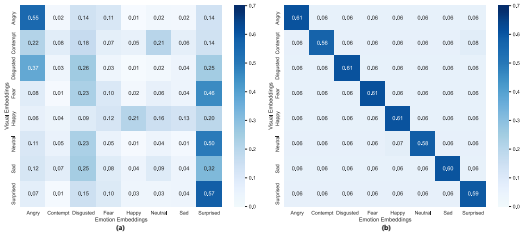

Fig. 7. **The heatmaps of the embedding distance between visual images and emotion-texts.** (a) The softmax values of cosine similarities between emotion-visual pairs before joint EVE processed. (b) The softmax values of cosine similarities between emotion-visual pairs after joint EVE processed.

F. Limitations

Despite the effective performance of our EAT-Face, we also find some limitations during explorations. Firstly, as shown in Fig. 6, there is a lower orientation for fear emotion that does not exceed 50%. The reason might be that there are similarities in facial movements (e.g. staring, wide mouth opening, etc.) between expressions of fear and surprise by the reconstructed characters, leading to conflicts. Secondly, issues of inter-frame jitter and mouth distortion persist in some cases. Moreover, our approach ignores the emotional information that speech audio itself may carry, which might lead to visual-audio emotional conflict. These will be part of our future work.

V. CONCLUSIONS AND FUTURE WORKS

A. Conclusions

In summary, we propose an emotion-controllable audio-driven talking face generation framework, called EAT-Face. In detail, to leverage multi-modal conditions, we design an LDM-based talking face reconstructor to synthesize audio-synced face images. Besides, we also propose a ControlNet-based facial emotion controller to manipulate emotional content. Under the introduction of joint emotion-visual embeddings, the semantic misalignment problem is effectually improved. Comprehensive experimental evaluations demonstrate the effectiveness of the proposed EAT-Face in generating high-fidelity and emotional talking face videos, which illustrates the proposed method is promising and has potential in future AIGC fields at the same time.

B. Future Works

Although our proposed method can manipulate facial emotions, in some circumstances the fusion of emotion embedding might lead to unexpected results such as strange mouth distortion, excessive and exaggerated expression, and so on, which can be further explored in subsequent works. Additionally, despite the usage of parallel sampling to speed up inference, it might cause a frame mutation that is not desired, which also can be a future direction for searching.

VI. ACKNOWLEDGEMENTS

This work is supported in part by National Natural Science Foundation of China (No.62025604, U2336208), and in part by Shenzhen Science and Technology Program (No. KQTD20221101093559018).

REFERENCES

[1] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, 33:12449–12460, 2020.

[2] D. Bigioi, S. Basak, H. Jordan, R. McDonnell, and P. Corcoran. Speech driven video editing via an audio-conditioned diffusion model. *arXiv preprint arXiv:2301.04474*, 2023.

[3] V. Blanz and T. Vetter. A morphable model for the synthesis of 3d faces. In *26th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH 1999)*, pages 187–194, 1999.

[4] H. Cao, D. G. Cooper, M. K. Keutmann, R. C. Gur, A. Nenkova, and R. Verma. Crema-d: Crowd-sourced emotional multimodal actors dataset. *IEEE Transactions on Affective Computing*, 5(4):377–390, 2014.

[5] L. Chen, G. Cui, C. Liu, Z. Li, Z. Kou, Y. Xu, and C. Xu. Talking-head generation with rhythmic head motion. In *European Conference on Computer Vision*, pages 35–51, 2020.

[6] J. S. Chung and A. Zisserman. Out of time: automated lip sync in the wild. In *Computer Vision–ACCV 2016 Workshops: ACCV 2016 International Workshops*, pages 251–263, 2017.

[7] J. Dai, R. Fan, Y. Song, Q. Guo, and F. He. Mean: An attention-based approach for 3d mesh shape classification. *The Visual Computer*, pages 1–14, 2023.

[8] J. Deng, J. Guo, N. Xue, and S. Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4690–4699, 2019.

[9] P. Dhariwal and A. Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021.

[10] C. Du, Q. Chen, T. He, X. Tan, X. Chen, K. Yu, S. Zhao, and J. Bian. Dae-talker: High fidelity speech-driven talking face generation with diffusion autoencoder. *arXiv preprint arXiv:2303.17550*, 2023.

[11] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. *Advances in Neural Information Processing Systems*, 27, 2014.

[12] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

[13] S. Goyal, S. Bhagat, S. Uppal, H. Jangra, Y. Yu, Y. Yin, and R. R. Shah. Emotionally enhanced talking face generation. In *Proceedings of the 1st International Workshop on Multimedia Content Generation and Evaluation: New Methods and Practice*, pages 81–90, 2023.

[14] J. Gu. Responsible generative ai: What to generate and what not. *arXiv preprint arXiv:2404.05783*, 2024.

[15] Y. Guo, K. Chen, S. Liang, Y.-J. Liu, H. Bao, and J. Zhang. Ad-nerf: Audio driven neural radiance fields for talking head synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5784–5794, 2021.

[16] J. Ho, W. Chan, C. Saharia, J. Whang, R. Gao, A. Gritsenko, D. P. Kingma, B. Poole, M. Norouzi, D. J. Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022.

[17] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

[18] J. Ho, C. Saharia, W. Chan, D. J. Fleet, M. Norouzi, and T. Salimans. Cascaded diffusion models for high fidelity image generation. *The Journal of Machine Learning Research*, 23(1):2249–2281, 2022.

[19] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

[20] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

[21] Y. Huang, F. Juefei-Xu, Q. Guo, Y. Liu, and G. Pu. Fakelocator: Robust localization of gan-based face manipulations. *IEEE Transactions on Information Forensics and Security*, 17:2657–2672, 2022.

[22] Y. Huang, F. Juefei-Xu, Q. Guo, Y. Liu, and G. Pu. Dodging deepfake detection via implicit spatial-domain notch filtering. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.

[23] Y. Huang, F. Juefei-Xu, R. Wang, Q. Guo, L. Ma, X. Xie, J. Li, W. Miao, Y. Liu, and G. Pu. Fakepolisher: Making deepfakes more detection-evasive by shallow reconstruction. In *Proceedings of the 28th ACM international conference on multimedia*, pages 1217–1226, 2020.

[24] Z. Huang, T. Zhang, W. Heng, B. Shi, and S. Zhou. Real-time intermediate flow estimation for video frame interpolation. In *European Conference on Computer Vision*, pages 624–642, 2022.

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

[26] X. Ji, H. Zhou, K. Wang, W. Wu, C. C. Loy, X. Cao, and F. Xu. Audio-driven emotional video portraits. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14080–14089, 2021.

[27] X. Jia, Y. Zhang, X. Wei, B. Wu, K. Ma, J. Wang, and X. Cao. Improving fast adversarial training with prior-guided knowledge. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[28] T. Karras, T. Aila, S. Laine, A. Herva, and J. 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.

[29] T. Karras, M. Aittala, T. Aila, and S. Laine. Elucidating the design space of diffusion-based generative models. *Advances in Neural Information Processing Systems*, 35:26565–26577, 2022.

[30] P. KR, R. Mukhopadhyay, J. Philip, A. Jha, V. Namboodiri, and C. Jawahar. Towards automatic face-to-face translation. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 1428–1436, 2019.

[31] A. Lahiri, V. Kwatra, C. Frueh, J. Lewis, and C. Bregler. 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*, pages 2755–2764, 2021.

[32] C. Lu, Y. Zhou, F. Bao, J. Chen, C. Li, and J. Zhu. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. *Advances in Neural Information Processing Systems*, 35:5775–5787, 2022.

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

[34] C. Lugaresi, J. Tang, H. Nash, C. McClanahan, E. Uboweja, M. Hays, F. Zhang, C.-L. Chang, M. G. Yong, J. Lee, et al. Mediapipe: A framework for building perception pipelines. *arXiv preprint arXiv:1906.08172*, 2019.

[35] A. Lugmayr, M. Danelljan, A. Romero, F. Yu, R. Timofte, and L. Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11461–11471, 2022.

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

[37] L. R. Medsker and L. Jain. Recurrent neural networks. *Design and Applications*, 5(64-67):2, 2001.

[38] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021.

[39] C. Mou, X. Wang, L. Xie, J. Zhang, Z. Qi, Y. Shan, and X. Qie. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. *arXiv preprint arXiv:2302.08453*, 2023.

[40] A. Nichol, P. Dhariwal, A. Ramesh, P. Shyam, P. Mishkin, B. McGrew, I. Sutskever, and M. Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021.

[41] A. Q. Nichol and P. Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, pages 8162–8171, 2021.

[42] P. Pataranutaporn, V. Danry, J. Leong, P. Punpongsanon, D. Novy, P. Maes, and M. Sra. Ai-generated characters for supporting personalized learning and well-being. *Nature Machine Intelligence*, 3(12):1013–1022, 2021.

[43] C. Qi, X. Cun, Y. Zhang, C. Lei, X. Wang, Y. Shan, and Q. Chen. Fatezero: Fusing attentions for zero-shot text-based video editing. *arXiv preprint arXiv:2303.09535*, 2023.

[44] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763, 2021.

[45] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.

[46] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. In

Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 10684–10695, 2022.

[47] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference*, pages 234–241, 2015.

[48] N. Ruiz, Y. Li, V. Jampani, Y. Pritch, M. Rubinstein, and K. Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22500–22510, 2023.

[49] C. Saharia, W. Chan, H. Chang, C. Lee, J. Ho, T. Salimans, D. Fleet, and M. Norouzi. Palette: Image-to-image diffusion models. In *ACM SIGGRAPH 2022 Conference Proceedings*, pages 1–10, 2022.

[50] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. L. Denton, K. Ghasemipour, R. Gontijo Lopes, B. Karagol Ayan, T. Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022.

[51] C. Saharia, J. Ho, W. Chan, T. Salimans, D. J. Fleet, and M. Norouzi. Image super-resolution via iterative refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4713–4726, 2022.

[52] S. Shen, W. Li, Z. Zhu, Y. Duan, J. Zhou, and J. Lu. Learning dynamic facial radiance fields for few-shot talking head synthesis. In *European Conference on Computer Vision*, pages 666–682, 2022.

[53] S. Shen, W. Zhao, Z. Meng, W. Li, Z. Zhu, J. Zhou, and J. Lu. Difftalk: Crafting diffusion models for generalized talking head synthesis. *arXiv preprint arXiv:2301.03786*, 2023.

[54] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[55] U. Singer, A. Polyak, T. Hayes, X. Yin, J. An, S. Zhang, Q. Hu, H. Yang, O. Ashual, O. Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022.

[56] J. Song, C. Meng, and S. Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

[57] Y. Song and S. Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in Neural Information Processing Systems*, 32, 2019.

[58] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

[59] M. Stypułkowski, K. Vougioukas, S. He, M. Zięba, S. Petridis, and M. Pantic. Diffused heads: Diffusion models beat gans on talking-face generation. *arXiv preprint arXiv:2301.03396*, 2023.

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

[61] J. Thies, M. Elgharib, A. Tewari, C. Theobalt, and M. Nießner. Neural voice puppetry: Audio-driven facial reenactment. In *Computer Vision–ECCV 2020: 16th European Conference*, pages 716–731, 2020.

[62] A. Van Den Oord, O. Vinyals, et al. Neural discrete representation learning. *Advances in Neural Information Processing Systems*, 30, 2017.

[63] L. Van der Maaten and G. Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(11), 2008.

[64] J. Wang, Y. Zhao, L. Liu, T. Xu, Q. Li, and S. Li. Emotional talking head generation based on memory-sharing and attention-augmented networks. *arXiv preprint arXiv:2306.03594*, 2023.

[65] K. Wang, Q. Wu, L. Song, Z. Yang, W. Wu, C. Qian, R. He, Y. Qiao, and C. C. Loy. Mead: A large-scale audio-visual dataset for emotional talking-face generation. In *European Conference on Computer Vision*, pages 700–717, 2020.

[66] T.-C. Wang, A. Mallya, and M.-Y. Liu. One-shot free-view neural talking-head synthesis for video conferencing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10039–10049, 2021.

[67] T. Xie, L. Liao, C. Bi, B. Tang, X. Yin, J. Yang, M. Wang, J. Yao, Y. Zhang, and Z. Ma. Towards realistic visual dubbing with heterogeneous sources. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 1739–1747, 2021.

[68] F. Yin, Y. Zhang, X. Cun, M. Cao, Y. Fan, X. Wang, Q. Bai, B. Wu, J. Wang, and Y. Yang. Styleheat: One-shot high-resolution editable talking face generation via pre-trained stylegan. In *European Conference on Computer Vision*, pages 85–101, 2022.

[69] C. Zhang, Y. Zhao, Y. Huang, M. Zeng, S. Ni, M. Budagavi, and X. Guo. Facial: Synthesizing dynamic talking face with implicit attribute learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3867–3876, 2021.

[70] L. Zhang and M. Agrawala. Adding conditional control to text-to-image diffusion models. *arXiv preprint arXiv:2302.05543*, 2023.

[71] X. Zhang, C. Zhang, T. Li, Y. Huang, X. Jia, X. Xie, Y. Liu, and C. Shen. A mutation-based method for multi-modal jailbreaking attack detection. *arXiv preprint arXiv:2312.10766*, 2023.

[72] W. Zhong, C. Fang, Y. Cai, P. Wei, G. Zhao, L. Lin, and G. 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.

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