EmotiveTalk: Expressive Talking Head Generation through Audio Information Decoupling and Emotional Video Diffusion

Haotian Wang[1], Yuzhe Weng[1], Yueyan Li[2], Zilu Guo[1], Jun Du[1*], Shutong Niu[1], Jiefeng Ma[1],
Shan He[3], Xiaoyan Wu[3], Qiming Hu[3], Bing Yin[3], Cong Liu[3], Qingfeng Liu[1]
[1] University of Science and Technology of China [2] Imperial College London [3] iFLYTEK
*Corresponding author : jundu@ustc.edu.cn

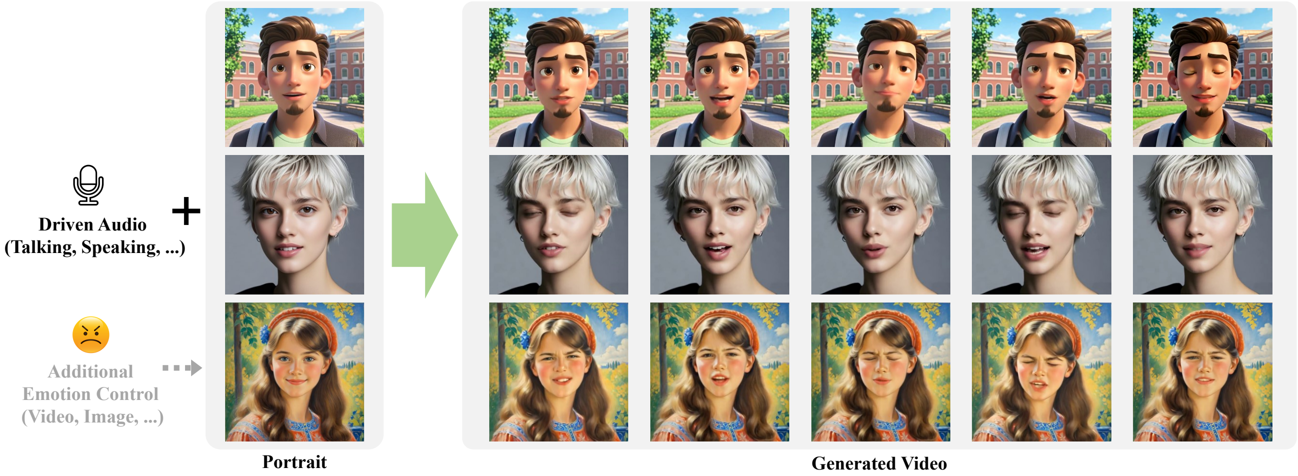


Figure 1. We propose **EmotiveTalk**, an expressive talking head generation framework. Taking a single portrait and the driven audio as input, our method can generate expressive portrait video sync with audio and customize the speaking style with additional emotion control.

Abstract

Diffusion models have revolutionized the field of talking head generation, yet still face challenges in expressiveness, controllability, and stability in long-time generation. In this research, we propose an EmotiveTalk framework to address these issues. Firstly, to realize better control over the generation of lip movement and facial expression, a Vision-guided Audio Information Decoupling (V-AID) approach is designed to generate audio-based decoupled representations aligned with lip movements and expression. Specifically, to achieve alignment between audio and facial expression representation spaces, we present a Diffusion-based Co-speech Temporal Expansion (Di-CTE) module within V-AID to generate expression-related representations under multi-source emotion condition constraints. Then we propose a well-designed Emotional Talking Head Diffusion (ETHD) backbone to efficiently generate highly expressive talking head videos, which contains an Expression Decoupling Injection (EDI) module to automatically decouple the expressions from reference portraits while integrating the target expression information, achieving more expressive generation performance. Experimental results show that EmotiveTalk can generate expressive talking head videos, ensuring the promised controllability of emotions and metric stability during long-time generation, yielding state-of-the-art performance compared to existing methods. The main page of our paper can be found in https://emotivetalk.github.io/.

1. Introduction

Talking head generation, also known as portrait image animation [52], demonstrates significant value across multiple domains, including television and film production, online education as well as human-machine interaction. The generation of realistic talking head videos involves two aspects of requirements. On the one hand, for the verbal aspect, it is essential to ensure the synchronization between speech and lip motions in the generated video [27]. On the other hand, for the non-verbal aspect, the generated video must convey non-verbal information, including facial expressions [26].

Despite the success of diffusion models [16, 24, 35] in image and video generation tasks, their application in talking head generation [15, 23, 31, 38, 45] still faces several challenges. For example, current methodologies [23, 38, 45] exhibit shortcomings in control of the generated emotional facial expressions, although they have made notable advancements in achieving synchronization between speech and lip movements. These audio-driven methods mainly directly synthesize expressions under weak audio conditions [38, 45]. However, the coupling of multiple information embedded in audio limits the effective learning of the mapping between speech and expressions and the controllability of generated emotion. Moreover, current diffusion-based methods often struggle to generate high-resolution video due to their large scale of parameters and the associated training costs [38, 45]. They also face challenges in stability during long-time generation due to their autoregressive inference strategies [23, 38, 45], which can lead to error accumulation across multiple inference clips.

To address these challenges, in this paper, we introduce EmotiveTalk, a highly expressive talking head generation framework with emotion control based on video diffusion. We propose a Vision-guided Audio Information Decouple (V-AID) approach to facilitate the decoupling of lip and expression related information contained in audio signals and also the alignment of audio representations with video representations under the guidance of vision facial motion information. Specifically, to achieve better alignment between speech and expression representation spaces, we present a Diffusion-based Co-speech Temporal Expansion (Di-CTE) module, which generates temporal expression-related representations from audio under utterance emotional conditions from multiple optional driven sources. Then, to effectively drive the decoupled representations, we propose an efficient video diffusion framework for expressive talking head generation that demonstrates effectiveness and enhanced stability in talking head video generation performance. The backbone incorporates an Expression Decoupling Injector (EDI) module in our backbone to achieve the automatic decoupling of expression information from the reference portrait while facilitating the injection of expression-driven information. In summary, our contributions are as follows: (1) We propose a Vision-Guided Audio Information Decouple (V-AID) approach that generates efficient decoupled lip-related and expression-related representations from audio for talking head generation. (2) We propose an Emotional Talking Head Diffusion (ETHD) framework that is capable of generating dynamic-length videos, which achieves highly expressive talking head video generation performance while ensuring metric stability over extended durations. (3) We further enhance emotion controllability by integrating conditions from emotion-driven sources and realizing the customization of generated emotions by multi-source emotion control.

2. Related Work

2.1. Audio-driven Talking Head Video Generation

The initial focus of the audio-driven talking head video generation task was on achieving synchronization between lip movements and the audio signal [27, 31]. Audio2Head [43] and SadTalker [49] integrate 3D information and control modules to enhance the naturalism of head movements. DreamTalk [23], Diffused Heads [36], and VASA-1 [46] further achieve more vivid and expressive results. Recently, a major shift occurred with the introduction of text-to-image pre-trained models. EMO [38], Hallo [45], and other similar frameworks [40, 44] built on the foundation of pre-trained image diffusion models [28] achieve high-fidelity talking head video generation results. Traditional audio-driven methods simply based on a data-driven approach, lack optional control on expression styles. Our model incorporates a decoupling mechanism that enables emotion control beyond conventional audio-driven frameworks.

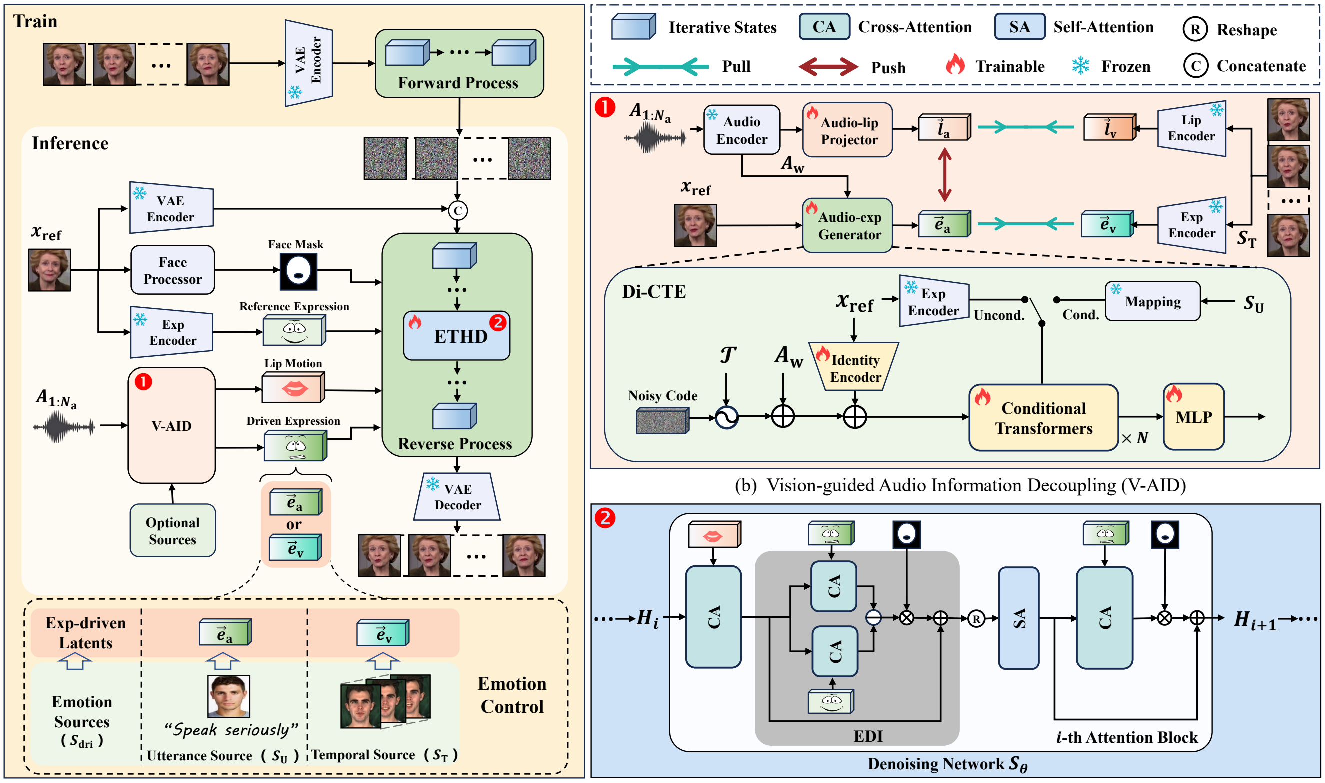
2.2. Controllable Talking Head Generation

Controlling the expression style in talking head video generation has long been a compelling challenge. Early methods [9, 12, 14, 20, 33, 37] model expressions in discrete emotion states, while recent methods [21–23, 41, 46] focus on transferring the expressions from a reference video to the generated video. Extracting decoupled representations of expressions is crucial for emotion transferring. Earlier approaches [22, 23] use 3DMM coefficients [4, 11] from reference videos, but this led to identity leakage issues, as the 3DMM coefficients encode not only expression information but also the speaker's facial structure information. PD-FGC [41] and AniTalker [21] employ contrastive learning approaches to acquire expression-related latent and realize expression driven with minor identity leakage.

In practical applications, emotion control information can originate from many other sources [1, 29]. In our approach, we derive a unified emotional control latent from various optional sources of emotion information and enable emotion control based on the emotion control latent.

2.3. Video Diffusion Models

The groundbreaking work on video diffusion is Video Diffusion Models (VDM) [18]. ImagenVideo [17] enhances VDM with cascaded diffusion models. Make-A-Video [32] and MagicVideo [53] then extend these concepts to enable seamless text-to-video transformations. AnimateDiff [13] utilizes a motion module to realize the conversion from text-to-image to text-to-video. Stable Video Diffusion (SVD) [5] implements innovative training strategies to generate high-fidelity videos. Our research utilizes diffusion models in expression-related latent generation and talking head rendering under facial motion control conditions.



(a) The training and inference pipeline of EmotiveTalk

(b) Vision-guided Audio Information Decoupling (V-AID)

(c) Emotional Talking Head Diffusion (ETHD) with Expression Decouple Injector (EDI)

Figure 2. The framework of EmotiveTalk. During the training process, the Vision-guided Audio Information Decouple (V-AID) module with Diffusion-based Co-speech Temporal Expansion (Di-CTE) expression generator in (b) is firstly trained to provide lip-related and expression-related representation from audio. Then the Emotional Talking Head Diffusion (ETHD) framework with Expression Decouple Injector (EDI) in (c) is trained with reference portrait condition and facial motion conditions to reconstruct the target frames, including lip-related and emotion-driven representation randomly chosen between $\vec{e}_a$ and $\vec{e}_v$ from V-AID module. During the inference process, EmotiveTalk takes portrait and speech audio as input, supplemented with optional emotion source S_{dri} to achieve emotion control.

3. Method

As shown in Fig. 2, the structure of the EmotiveTalk is divided into two main parts: (1) the Vision-guided Audio Information Decouple (V-AID) with Diffusion-based Co-speech Temporal Expansion (Di-CTE) module; (2) the Emotional Talking Head Diffusion (ETHD) framework with Expression Decoupling Injector (EDI) module.

3.1. Preliminary

Task Definition. The task of controllable talking head generation involves creating a vivid talking head video from two inputs: a static single-person portrait $\boldsymbol{x}_{\text{ref}}$, and a driven speech sequence $\boldsymbol{A} \in \mathbb{R}^{N_a}$. Besides, emotion sources $\boldsymbol{S}_{\text{dri}}$ can also utilized as optional input to realize better controllability of emotion. When the optional $\boldsymbol{S}_{\text{dri}}$ is not provided, our method aims to generate expression-related representation solely by the speech input and the portrait $\boldsymbol{x}_{\text{ref}}$. The output is the generated video frames $\hat{\boldsymbol{X}}_{1:N} = \{\hat{\boldsymbol{x}}_0, ..., \hat{\boldsymbol{x}}_N\}$.

Diffusion Models. Let $\boldsymbol{X}_{(0)}$ represent video latents sampled from a given distribution $q(\boldsymbol{X}_{(0)})$. In the forward diffusion process, Gaussian noise is progressively added to $\boldsymbol{X}_0$, gradually diffusing towards a distribution resembling $\mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$. This process forms a fixed Markov chain [24, 35]:

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

where $\{\beta_t\}_{t=1}^{T}$ are known constants. Notably, the marginal distribution at any time can directly derive from $\boldsymbol{X}_{(0)}$ as:

$$q(\boldsymbol{X}_{(t)}|\boldsymbol{X}_{(0)}) = \mathcal{N}(\boldsymbol{X}_{(t)}; \sqrt{\bar{\alpha}_t}\,\boldsymbol{X}_{(0)}, (1-\bar{\alpha}_t)\boldsymbol{I}) \quad (2)$$

where $\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i$ and $\alpha_t = 1 - \beta_t$. The reverse process gradually recovers the original video latent from the noisy latent $\boldsymbol{X}_{(T)} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$, achieving by training a network to predict the posterior distribution $p_\theta(\boldsymbol{X}_{(t-1)}|\boldsymbol{X}_{(t)}, \boldsymbol{c})$ under condition set $\boldsymbol{c}$. To learn $p_\theta(\boldsymbol{X}_{(t-1)}|\boldsymbol{X}_{(t)})$. The model is trained using the following loss function:

$$L = \mathbb{E}_{t,\boldsymbol{X}_0,\epsilon,\boldsymbol{c}}[\|\epsilon - \epsilon_\theta(\boldsymbol{X}_{(t)}, t, \boldsymbol{c})\|^2] \quad (3)$$

We use the diffusion strategy for expression-related representation generation in Di-CTE and video latent generation.

3.2. Vision-guided Audio Information Decouple

Speech is rich in plentiful coupled information, previous methods focused on decoupling speech information in the audio space [47, 54]. However, the representations obtained through these approaches are generally not well-suited for talking head generation, due to the inherent disparity between the audio and facial motion representations. We propose that facial motion information in the vision space can guide the decoupling of coupled speech information due to the correlation between speech information and different facial motions [48] and also facilitate the generation of aligned facial motion related representations from audio. Based on this, we designed a Vision-guided Audio Information Decoupling (V-AID) module. This module takes audio sequence $\boldsymbol{A}$ and reference portrait $\boldsymbol{x}_{\text{ref}}$ as input. The audio stream first passes through a pre-trained Wav2Vec audio encoder [2], followed by the trainable audio-to-lip projector and audio-to-expression generator to obtain lip and expression-related latents. The two modules are trained under the supervision of lip and expression representations of vision space, elaborated in the supplementary material.

Audio-lip Contrastive Learning. We leverage the latent representation of lip motions in vision space to guide the audio-to-lip mapping, thereby achieving alignment between the audio and lip motion representations. Specifically, we use a pre-trained lip encoder to extract decoupled lip-related latents $\vec{\boldsymbol{l}}_{\text{v}} = \{l_1, ..., l_N\}$ from videos paired with audio. The audio stream is processed through an audio-to-lip projector with a Perceiver Transformer [19] architecture detailed in the supplementary material to generate lip-related latents $\vec{\boldsymbol{l}}_{\text{a}} = \{\hat{l}_1, ..., \hat{l}_N\}$. The infoNCE [25] contrastive loss function is utilized to optimize the lower bound of mutual information (MI) between $\vec{\boldsymbol{l}}_{\text{a}}$ and $\vec{\boldsymbol{l}}_{\text{v}}$ to maximize MI between frame-level lip movements and the corresponding driving speech signal, where $(\hat{l}_i, l_i)$ denotes a positive pair and $(\hat{l}_i, l_j)$ denotes negative pairs. The loss function is formulated as follows, with $\text{sim}(\cdot)$ represents cosine similarity:

$$\mathcal{L}_{\text{lipc}} = -\frac{1}{N}\sum_{i=1}^{N}\log\left(\frac{\exp\left(\frac{\text{sim}(\hat{l}_i, l_i)}{\tau}\right)}{\sum_{j=1}^{N}\exp\left(\frac{\text{sim}(\hat{l}_i, l_i)}{\tau}\right)}\right) \tag{4}$$

Furthermore, we also supplement the contrastive learning loss with Mean Squared Error (MSE) loss to synchronize both the motion and morphological information between $\vec{\boldsymbol{l}}_{\text{a}}$ and $\vec{\boldsymbol{l}}_{\text{v}}$. The loss function is as follows:

$$\mathcal{L}_{\text{lipm}} = \frac{1}{N}\sum_{i=1}^{N}||l_i - \hat{l}_i||^2 \tag{5}$$

The final training loss function is the combination of two losses, as follows:

$$\mathcal{L}_{\text{lip}} = \alpha\mathcal{L}_{\text{lipc}} + \beta\mathcal{L}_{\text{lipm}} \tag{6}$$

Di-CTE for Audio-to-expression Generation. We utilize representations of facial expressions from the vision space to guide the alignment of audio-based emotion information with facial expressions. Generally, speech and facial expressions are not strictly correlated on a one-to-one basis, the same speech can correspond to different but plausible facial expressions. To address this, we propose a Diffusion-based Co-speech Temporal Expansion (Di-CTE) module to generate frame-level expression-related latent $\vec{\boldsymbol{e}}_{\text{a}}$ from initial expression under speech constraints, leveraging the advantages of diffusion models in terms of generative diversity. We leverage a pre-trained expression encoder to extract decoupled expression latent $\vec{\boldsymbol{e}}_{\text{v}}$ from ground-truth video as vision supervision. Di-CTE inputs consist of a reference frame ($\boldsymbol{x}_{\text{ref}}$) from the ground truth video serving as speaker identity and speech embedding $\boldsymbol{A}_{\text{w}}$ to provide temporal emotion information. During training, the emotion condition $\boldsymbol{e}_{\text{cond}}$ is provided by the first frame of the ground-truth video, and the output is expression-related latent $\vec{\boldsymbol{e}}_{\text{a}}$ sync with the speech. The denoising loss of network S_θ is defined as follows, where t denotes the DDPM step:

$$\mathcal{L}_{\text{exp}} = ||\vec{\boldsymbol{e}}_{\text{v}} - S_\theta(\vec{\boldsymbol{e}}_{\text{a}(t)}, t, \boldsymbol{x}_{\text{ref}}, \boldsymbol{A}_{\text{w}}, \boldsymbol{e}_{\text{cond}})||^2 \tag{7}$$

Mutual Information Constraint. Finally, to decouple lip-related and expression-related information and mitigate their mutual interference, we introduced a mutual information (MI) constraint during the joint training of the audio-to-lip and audio-to-expression modules. Specifically, we employ CLUB [7] to optimize the upper bound of MI between the lip-related latent $\vec{\boldsymbol{l}}_{\text{a}}$ from the audio-to-lip module and the expression-related latent $\vec{\boldsymbol{e}}_{\text{a}}$ from the audio-to-expression module. The total loss function is as follows:

$$\mathcal{L}_{\text{V-AID}} = \mathcal{L}_{\text{lip}} + \mathcal{L}_{\text{exp}} + \text{CLUB}\{\vec{\boldsymbol{l}}_{\text{a}}, \vec{\boldsymbol{e}}_{\text{a}}\} \tag{8}$$

By minimizing the MI between $\vec{\boldsymbol{l}}_{\text{a}}$ and $\vec{\boldsymbol{e}}_{\text{a}}$, we achieve a separation of the two representation spaces.

3.3. Emotional Talking Head Diffusion

In this subsection, we present a diffusion-based framework for generating emotional talking heads. Before ETHD, the driving audio is processed through V-AID to obtain lip-related and expression-related latents, and the portrait is projected into latent space via temporal Variational Autoencoder (VAE) and concatenated with input noise along the channel dimension, shown in Fig. 2. ETHD outputs a sequence of frame latents synchronized with the speech.

Backbone Network. Our backbone network leverages a 3D-Unet architecture with the spatial-temporal separable attention mechanism [5]. The spatial attention module comprises two blocks. Firstly, the lip-related latent $\vec{\boldsymbol{l}}_{\text{a}}$ is injected through spatial cross attention. Then, an Expression Decoupling Injector (EDI) module, articulated late in Section 3.3, is employed to integrate expression-driven latent

$\vec{\boldsymbol{e}}_{\text{dri}}$ ($\vec{\boldsymbol{e}}_{\text{dri}} = \vec{\boldsymbol{e}}_{\text{a}}$ for audio-only driven task and $\vec{\boldsymbol{e}}_{\text{dri}} = \vec{\boldsymbol{e}}_{\text{v}}$ for video-driven task). Analogously, the temporal attention module also encompasses two components: a temporal self-attention mechanism and a temporal cross-attention module. The temporal cross-attention module engages in cross-attention with expression-driven latent to learn subtle temporal variations in emotional expression. The output latents are then processed through a temporal VAE decoder to obtain the generated motion frames.

Expression Decoupling Injector. In talking head generation, the inherent expression information in the reference portrait usually constrains the generation of the target expression, leading to sub-optimal expressive results. To address this, we propose an Expression Decoupling Injection (EDI) module to achieve emotional expressions by automatically decoupling the expression information from reference portraits while integrating the expression-driven information, which consists of two parallel attention branches. One branch computes the attention between the hidden states $\boldsymbol{H}_i \in \mathbb{R}^{f\times h\times w\times c}$ (f is the number of processed frames, h and w is the height and width of hidden states, c is the number of channels) and the expression embeddings $\vec{\boldsymbol{e}}_{\text{ref}}$ of the reference portrait while the other branch computes the attention between the hidden states $\boldsymbol{H}_i$ and the expression-driven representation $\vec{\boldsymbol{e}}_{\text{dri}}$. By subtracting these two cross-attention outputs, we achieve the transition of facial expressions in the generated video from the expression of the reference image to the driving expression, as shown in the following equation:

$$\boldsymbol{Attn}_i = \text{CrossAttn}(\boldsymbol{H}_i, \vec{\boldsymbol{e}}_{\text{dri}}) - \text{CrossAttn}(\boldsymbol{H}_i, \vec{\boldsymbol{e}}_{\text{ref}}) \quad (9)$$

Moreover, to enforce the expression-related latent act only on the facial region without affecting the lip region generation, we apply an attention mask similar to Hallo [45] to the resulting attention value. Specifically, we use the off-the-shelf toolbox OpenFace [3] to predict landmarks from portrait $\boldsymbol{x}_{\text{ref}}$ and calculate binary bounding box masks $\mathbf{M}_{\text{lip}}, \mathbf{M}_{\text{face}} \in \{0,1\}^{h\times w}$ which indicate the inner of lip region and face region. Then, the output of the EDI block is formulated based on bounding box masks, as follows:

$$\boldsymbol{H}_i^{\text{spa}} = \boldsymbol{H}_i + \boldsymbol{Attn}_i \odot (1 - \mathbf{M}_{\text{lip}}) \odot \mathbf{M}_{\text{face}} \quad (10)$$

Expression Temporal Cross-attention. To implement better modeling of the time-variance of facial expressions, we introduce a temporal cross-attention module. Specifically, we squeeze the spatial dimensions of the hidden states $\boldsymbol{H}_i^{\text{spa}}$ to $\boldsymbol{H}_i^{\text{tem}} \in \mathbb{R}^{(h\times w)\times f\times c}$ and compute the cross-attention between $\boldsymbol{H}_i^{\text{tem}}$ and the expression-driven latent $\vec{\boldsymbol{e}}_{\text{dri}}$. This makes the model more sensitive to the temporal correlations of emotional information. Additionally, the same bounding box masks are utilized to constrain the sensible area of attention calculation.

3.4. Training and Inference

Training. The V-AID module in Sec. 3.2 is first pretrained to generate decoupled lip-related representation $\vec{\boldsymbol{l}}_{\text{a}}$ and emotion-related representation $\vec{\boldsymbol{e}}_{\text{a}}$ from driven audio window A_{w} and then remain frozen while training the ETHD backbone.

Subsequently, we train the ETHD backbone by sampling tuples $(\boldsymbol{X}, \boldsymbol{x}_{\text{ref}}, t, \vec{\boldsymbol{l}}_{\text{a}}, \vec{\boldsymbol{e}}_{\text{ref}}, \vec{\boldsymbol{e}}_{\text{dri}})$, $\vec{\boldsymbol{e}}_{\text{dri}}$ is random choice in video expression-related representation $\vec{\boldsymbol{e}}_{\text{v}}$ and the generated expression-related representation $\vec{\boldsymbol{e}}_{\text{a}}$. The total denoising loss function is formulated as:

$$\mathcal{L}_{\text{de}} = ||\boldsymbol{X}_{(0)} - S_\theta(\boldsymbol{X}_{(t)}, \boldsymbol{x}_{\text{ref}}, t, \vec{\boldsymbol{l}}_{\text{a}}, \vec{\boldsymbol{e}}_{\text{ref}}, \vec{\boldsymbol{e}}_{\text{dri}})||^2 \quad (11)$$

Inference. In the inference phase, we employ a non-autoregressive inference method to avoid the accumulation of error. Specifically, when performing long-time generation, we sample a Gaussian-like noisy latent and divide the total duration into several overlapping clips with a defined window size. We utilize DDIM [35] sampler for ETHD to denoise each clip sequentially per step, then we assign a weighting strategy the same as MimicMotion [50] to assign higher fusion weights for frame latents closer to the center of each clip. Repeat this process iteratively to obtain the clean frame latent. This approach allows us to perform inference of arbitrary lengths without error accumulation.

3.5. Multi-source Emotion Control

To flexibly control emotional expression in generated video based on control sources, we designed the Multi-source Emotion Control (MEC) pipeline. MEC introduces time-varying facial expressions to the generated video based on optional temporal or utterance sources.

Temporal Sources Emotion Control. External expression-driven videos are treated as temporal sources, denoted as $\boldsymbol{S}_{\text{T}}$, due to their rich temporal variations in expression. We directly apply the pre-trained expression encoder to extract the expression-driven latent $\vec{\boldsymbol{e}}_{\text{v}}$, as detailed in Section 3.3. The final emotive video is rendered using exp-driven latent $\vec{\boldsymbol{e}}_{\text{v}}$, and lip-related latent $\vec{\boldsymbol{l}}_{\text{a}}$, derived from Section 3.2.

Utterance Sources Emotion Control. To improve the temporal dynamism and better alignment with the driving speech of generated expressions based on utterance sources $\boldsymbol{S}_{\text{U}}$ that only provide general emotional information $\boldsymbol{e}_{\text{cond}}$, we use the Di-CTE module (Section 3.2) to generate frame-level expression-driven latent $\vec{\boldsymbol{e}}_{\text{a}}$ from $\boldsymbol{e}_{\text{cond}}$. Specifically, for expression-driven images of different people ($\boldsymbol{x}_{\text{dri}}$), we map the image to the emotion condition latent space $\boldsymbol{e}_{\text{cond}}$ using pre-trained expression encoder (Section 3.2). For cross-modality control sources like $\boldsymbol{t}_{\text{dri}}$, we apply a cross-modality mapping to align with $\boldsymbol{e}_{\text{cond}}$, detailed in the supplementary material. The final emotive video is rendered using lip-related latent $\vec{\boldsymbol{l}}_{\text{a}}$ and expression-driven latent $\vec{\boldsymbol{e}}_{\text{a}}$ via our diffusion backbone, as detailed in Section 3.3.

Methods	HDTF / MEAD					
	Driven	FID (↓)	FVD (↓)	Sync-C (↑)	Sync-D (↓)	E-FID (↓)
SadTalker [49]	A	22.34 / **36.88**	589.63 / **132.27**	7.75 / 6.46	7.36 / 8.07	0.66 / 1.14
AniTalker [21]	A	51.66 / 68.01	583.70 / 941.49	7.73 / 6.76	7.43 / 7.64	1.11 / 1.11
AniPortrait [44]	A	17.71 / 42.43	676.30 / 379.08	3.75 / 2.30	10.63 / 12.38	1.21 / 2.69
Hallo [45]	A	17.15 / 52.07	276.31 / 210.56	7.99 / **7.45**	7.50 / 7.47	0.65 / 0.60
Ours	A	**16.64** / 53.21	**140.96** / 207.67	**8.24** / 6.82	**7.09 / 7.43**	**0.54 / 0.57**
PD-FGC [41]	A+V	67.97 / 121.46	464.90 / 353.75	7.30 / 5.15	7.72 / 8.77	0.74 / 1.92
StyleTalk [22]	A+V	29.65 / 118.48	184.60 / 197.18	4.34 / 3.86	10.35 / 10.74	0.42 / 0.56
DreamTalk [23]	A+V	29.37 / 105.92	263.78 / 204.48	6.80 / 5.64	8.03 / 8.69	0.55 / 0.87
Ours	A+V	**16.09 / 50.84**	**120.70 / 153.71**	**8.41 / 6.79**	**7.11 / 7.58**	**0.34 / 0.40**
Ground Truth	A+V	-	-	8.63 / 7.30	6.75 / 8.31	-

Table 1. Overall comparisons on HDTF and MEAD. "A" denotes audio-only driven and "A+V" denotes audio-video driven. "↑" indicates better performance with higher values, while "↓" indicates better performance with lower values.

4. Experiment

4.1. Experimental Setup

Implementation Details. Experiments encompassing both training and inference were carried out on open-source datasets HDTF [51] and MEAD [42], which consist of talking individuals videos of diverse genders, ages, and ethnicities. We utilize a two-stage training strategy, firstly, we trained the V-AID module with a learning rate of 1e-4. In the second stage, the audio-to-video diffusion backbone was trained while the pre-trained V-AID modules remained frozen in training. Notably, thanks to the efficient design of our model, we can conduct high-resolution and long-time video training. We conduct a training configuration of the resolution of 512×512 and 120 frames. The learning rate is set to 1e-5 with a batch size of 1. Our backbone also supports up to 1024×1024 training, and experiments on other configurations are detailed in the supplementary material.

During the inference, we use the sampling algorithm of DDIM [35] to generate the video clip for 25 steps, the inference window size is as same as the training frame number and the overlap is set to 1/5 of the window size.

Evaluation Metrics. The proposed framework has been evaluated with several quantitative metrics including Fréchet Inception Distance (FID) [30], Fréchet Video Distance (FVD) [34, 39], Synchronization-C (Sync-C) [8], Synchronization-D (Sync-D) [8] and E-FID [38]. Specifically, FID and FVD conduct the image-level and frame-level measurement of the quality of the generated frames and the similarity between generated and ground-truth frames, with lower values indicating better performance. The SyncNet scores assess the lip synchronization quality, with higher Sync-C and lower Sync-D scores indicating better alignment with the driven speech signal. Additionally, to evaluate the expressiveness of the facial expressions in the generated videos, we also utilize the Expression-FID (E-FID) metric introduced in EMO [38] to quantitatively measure the expression divergence between the synthesized videos and gound-truth videos.

Baselines. We conducted a comparative analysis of our proposed method against several open-source implementations, including audio-only driven strategies including SadTalker [49], AniPortrait [44], AniTalker [21] and Hallo [45], and audio-video driven strategies including PD-FGC [41], StyleTalk [22], and DreamTalk [23]. For audio-only driven comparison, our framework derives lip-related and expression-driven latents solely from the audio and reference portrait input. As for audio-video driven, the expression-driven latent is derived from the paired video.

4.2. Overall Evaluation

Tab. 1 shows the results of the comprehensive comparison with other methods. Overall, methods based on Stable Diffusion like Hallo achieve optimal FID scores, confirming the potential of diffusion models in generating high-fidelity videos. Also, audio-video driven methods perform better on the E-FID metric, benefiting from the inclusion of expression cues derived from video. Our method outperforms previous methods in both audio-only driven and video-driven tasks across most metrics, especially on E-FID and SyncNet metrics, highlighting its superior capabilities of generating high-fidelity and vivid videos. More comparison results can be found in the supplementary material.

4.3. Ablation Study

To analyze the contributions of our designs, we conduct ablation studies on our main modules.

Effectiveness of V-AID. We conduct an ablation study

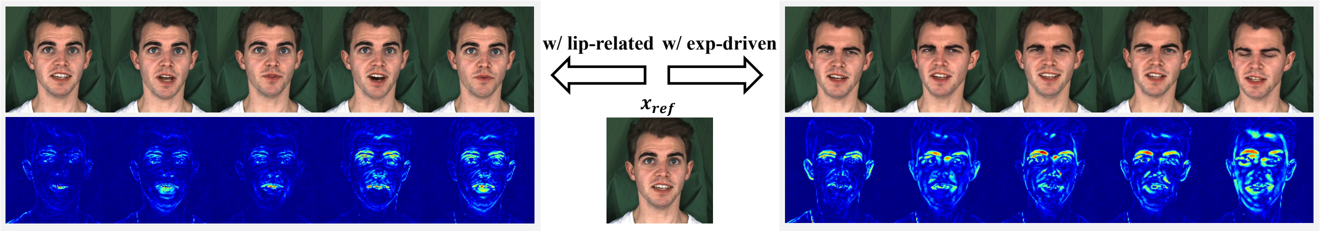


Figure 3. The visualize results of generated frames and difference heatmap with the reference portrait based on lip-related and expression-driven representations driven separately.

Methods	FID (↓)	Sync-C (↑)	E-FID (↓)
V-AID	16.64	**8.24**	**0.54**
w/o lip-related	16.02	0.65	0.57
w/o exp-driven	**14.86**	8.04	1.23
no decouple	16.98	7.72	0.66

Table 2. Ablation comparison on V-AID on HDTF dataset.

with three variants: (1) driven only by original audio embedding without V-AID (no decouple); (2) driven without lip-related latents (w/o lip-related); (3) driven without expression-driven latents (w/o exp-driven). Our full model is denoted as (V-AID). The experiment is carried out in the test subset of HDTF. Shown in Tab 2, the results indicate that using V-AID shows improvements across all three metrics compared to direct injection without decoupling, with notable gains in the Sync-C and E-FID metrics. Additionally, we observe a significant drop in Sync-C when lip-related latents are removed, and a substantial degradation in E-FID when expression-driven latents are excluded. This supports the different roles that the two representations play in driving lip movement and facial expressions. Furthermore, we observe that FID achieves the best performance without expression-driven, which is due to the higher similarity between generated frames and reference images when expression-driven latents are excluded, further confirmed in subsequent experiments. More detailed quality ablation results of V-AID are provided in the supplementary materials.

Effectiveness of Decoupled Representations. To evaluate the decoupling ability of two representations, we utilized the lip-related and expression-driven latents from V-AID to generate videos separately and visualize the results. Shown in Fig. 3, the results indicate that the main movement occurs at the lip region of the generated frames driven by the lip-related latents. In contrast, the generated frames driven by the expression-driven latents exhibit substantial changes in facial expressions compared to the reference portrait, with higher heat values distributed across the entire facial area, particularly in the eye region. The results demonstrate the effectiveness of decoupled lip and expression representations in controlling facial motions separately.

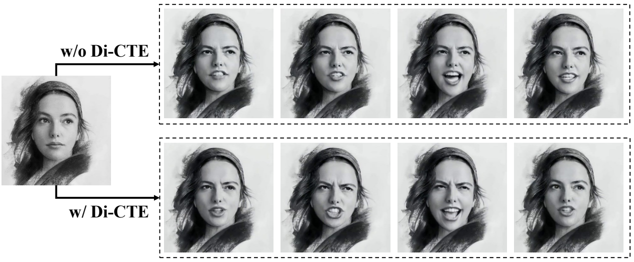


Figure 4. Results on expression generation w/ or w/o Di-CTE.

Length	FID (↓)	Sync-C (↑)	E-FID (↓)
120frames	16.78	8.25	0.60
250frames	16.96	8.21	0.67
750frames	16.93	8.46	0.62
1500frames	16.97	8.40	0.61

Table 3. Comparison on long-time generation on HDTF dataset.

Effectiveness of Di-CTE. To validate the superiority of our proposed Di-CTE module in expanding utterance driven sources to generate time-variance expressions, we employed a single expression-driven image and conducted inference using two configurations: with the Di-CTE module (w/ Di-CTE) and without the Di-CTE module (w/o Di-CTE). The facial expressions in inference results w/o Di-CTE module activated show minimal temporal variation in Fig. 4, while more expressive and vivid results are achieved by the Di-CTE activated, demonstrating its effectiveness.

Effectiveness on Long-time Generation. To validate the stability of long-time generation, we conducted studies on generating with varying lengths by audio-only driven. We employed four different test configurations, ranging from short to long duration, and evaluated identity consistency, lip-sync accuracy, and expression alignment across varying generation durations. The results are presented in Tab 3.

The results indicate that as inference duration increases, the FID, SyncNet, and E-FID metrics exhibit relatively minor fluctuations without degradation trend over time, confirming the stability of EmotiveTalk in long-time inference scenarios.

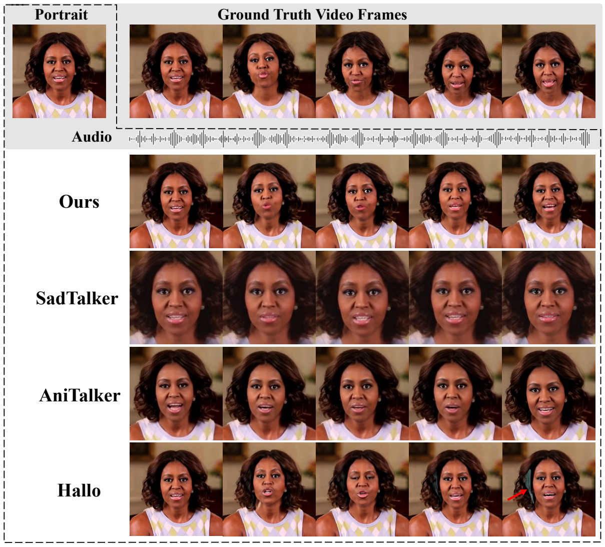


Figure 5. Case study on audio-only driven approaches.

Methods	Lip-Sync (↑)	Exp-Q (↑)	Realness (↑)	V-Q (↑)
SadTalker [49]	3.03	3.03	3.01	3.29
AniTalker [21]	2.82	3.04	2.87	3.24
AniPortrait [44]	1.65	1.79	1.65	2.26
Hallo [45]	3.73	3.36	3.28	3.49
StyleTalk [22]	2.50	2.88	2.78	3.02
DreamTalk [23]	3.69	3.45	3.40	3.38
Ours	**4.15**	**3.96**	**3.98**	**4.03**
Ground Truth	4.51	4.49	4.44	4.40

Table 4. User Study Results.

4.4. Case Study

Comparison on Audio-only Driven. Fig. 5 shows the qualitative results on audio-only driven approaches. The results show that AniTalker and SadTalker struggle to generate video faithful to the reference image x_{ref} due to the cropping and warping operation and also fall short in lip synchronization. Hallo demonstrates the ability to preserve speaker identity, but encounters instability issues in video generation, resulting in the unintended appearance of artifacts. Our method surpasses previous approaches in achieving lip synchronization, identity maintenance, and generation stability, resulting in the best overall performance.

Comparison on Emotion Control. To evaluate the performance of emotion control, we use a portrait paired with a happy video from another person and employ various methods to transfer the emotion. Fig. 6 shows the results, which indicate that StyleTalk and DreamTalk struggle in lip synchronization due to the coupling of lip and expression. PD-FGC faces the challenge of lip shape deformation. Our method achieves the most neutral and expressive emotion control results also ensures lip sync, highlighting the effectiveness of our decoupling approach and model design.

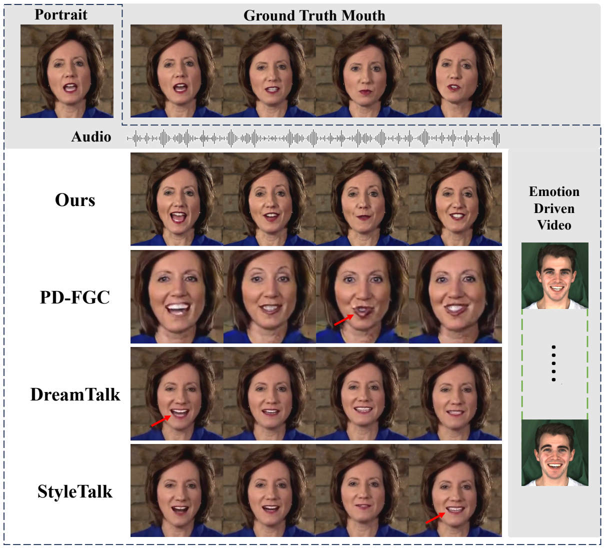


Figure 6. Case study on emotion control approaches.

4.5. User Study

We generated 10 test samples covering various emotion states and used 7 different models to generate with the ground-truth samples included. We conducted a user study of 26 participants, for each method, the participant is required to score 10 videos sampled from the test samples and is asked to give a rating (from 1 to 5, 5 is the best) on four aspects: (1) the lip sync quality (Lip-Sync), (2) the quality of expressions (Exp-Q), (3) the realness of results (Realness), (4) the quality of generated video (V-Q). The results are shown in Tab. 4, our method outperforms existing approaches across all aspects, particularly in expression quality and lip sync, highlighting its superior capabilities.

5. Conclusion

In this work, we propose EmotiveTalk, a novel method that aims at enhancing the emotional expressiveness and controllability of talking head video generation. We propose a novel approach to decouple audio embedding by leveraging facial motion information, enabling the generation of decoupled representations that correspond directly to lip motions and facial expressions. Additionally, we introduce a well-designed video diffusion framework that drives these representations to generate expressive talking head videos. We further enhance the emotion control ability by incorporating additional emotion information from multiple sources to customize the generated emotions. Extensive experiments demonstrate the superiority of EmotiveTalk.

Acknowledgments. This work was supported by the National Natural Science Foundation of China under Grant 62171427 and the National Key R&D Program of China under Grant 2022YFB4500600.

References

[1] Naveed Ahmed, Zaher Al Aghbari, and Shini Girija. A systematic survey on multimodal emotion recognition using learning algorithms. *Intelligent Systems with Applications*, 2023. 2

[2] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460, 2020. 4, 2

[3] Tadas Baltrušaitis, Peter Robinson, and Louis-Philippe Morency. Openface: an open source facial behavior analysis toolkit. In *2016 IEEE winter conference on applications of computer vision (WACV)*, pages 1–10. IEEE, 2016. 5

[4] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3d faces. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pages 157–164. 2023. 2

[5] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023. 2, 4, 3

[6] Zhiyuan Chen, Jiajiong Cao, Zhiquan Chen, Yuming Li, and Chenguang Ma. Echomimic: Lifelike audio-driven portrait animations through editable landmark conditions. *arXiv preprint arXiv:2407.08136*, 2024. 4

[7] Pengyu Cheng, Weituo Hao, Shuyang Dai, Jiachang Liu, Zhe Gan, and Lawrence Carin. Club: A contrastive log-ratio upper bound of mutual information. In *International conference on machine learning*. PMLR, 2020. 4

[8] Joon Son Chung and Andrew Zisserman. Out of time: automated lip sync in the wild. In *Computer Vision–ACCV 2016 Workshops: ACCV 2016 International Workshops, Taipei, Taiwan, November 20-24, 2016, Revised Selected Papers, Part II 13*, pages 251–263. Springer, 2017. 6

[9] Radek Daněček, Kiran Chhatre, Shashank Tripathi, Yandong Wen, Michael Black, and Timo Bolkart. Emotional speech-driven animation with content-emotion disentanglement. In *SIGGRAPH Asia 2023 Conference Papers*, 2023. 2

[10] Nikita Drobyshev, Antoni Bigata Casademunt, Konstantinos Vougioukas, Zoe Landgraf, Stavros Petridis, and Maja Pantic. Emoportraits: Emotion-enhanced multimodal one-shot head avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 4

[11] Bernhard Egger, William AP Smith, Ayush Tewari, Stefanie Wuhrer, Michael Zollhoefer, Thabo Beeler, Florian Bernard, Timo Bolkart, Adam Kortylewski, Sami Romdhani, et al. 3d morphable face models—past, present, and future. *ACM Transactions on Graphics (ToG)*, 39(5):1–38, 2020. 2

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

[13] Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023. 2

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

[15] Tianyu He, Junliang Guo, Runyi Yu, Yuchi Wang, Jialiang Zhu, Kaikai An, Leyi Li, Xu Tan, Chunyu Wang, Han Hu, et al. Gaia: Zero-shot talking avatar generation. *arXiv preprint arXiv:2311.15230*, 2023. 2

[16] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 1

[17] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022. 2

[18] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *Advances in Neural Information Processing Systems*, 35:8633–8646, 2022. 2

[19] Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. Perceiver: General perception with iterative attention. In *International conference on machine learning*, pages 4651–4664, 2021. 4, 2

[20] Xinya Ji, Hang Zhou, Kaisiyuan Wang, Wayne Wu, Chen Change Loy, Xun Cao, and Feng Xu. Audio-driven emotional video portraits. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14080–14089, 2021. 2

[21] Tao Liu, Feilong Chen, Shuai Fan, Chenpeng Du, Qi Chen, Xie Chen, and Kai Yu. Anitalker: Animate vivid and diverse talking faces through identity-decoupled facial motion encoding. *arXiv preprint arXiv:2405.03121*, 2024. 2, 6, 8

[22] Yifeng Ma, Suzhen Wang, Zhipeng Hu, Changjie Fan, Tangjie Lv, Yu Ding, Zhidong Deng, and Xin Yu. Styletalk: One-shot talking head generation with controllable speaking styles. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1896–1904, 2023. 2, 6, 8, 4

[23] Yifeng Ma, Shiwei Zhang, Jiayu Wang, Xiang Wang, Yingya Zhang, and Zhidong Deng. Dreamtalk: When expressive talking head generation meets diffusion probabilistic models. *arXiv preprint arXiv:2312.09767*, 2023. 2, 6, 8, 4

[24] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International conference on machine learning*. PMLR, 2021. 1, 3

[25] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 4

[26] Deepika Phutela. The importance of non-verbal communication. *IUP Journal of Soft Skills*, 9(4):43, 2015. 1

[27] KR Prajwal, Rudrabha Mukhopadhyay, Vinay P Namboodiri, and CV Jawahar. A lip sync expert is all you need for speech to lip generation in the wild. In *Proceedings of the*

28th ACM international conference on multimedia, 2020. 1, 2

[28] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn 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. 2, 3, 6

[29] Anvita Saxena, Ashish Khanna, and Deepak Gupta. Emotion recognition and detection methods: A comprehensive survey. *Journal of Artificial Intelligence and Systems*, 2020. 2

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

[31] Shuai Shen, Wenliang Zhao, Zibin Meng, Wanhua Li, Zheng Zhu, Jie Zhou, and Jiwen Lu. Difftalk: Crafting diffusion models for generalized audio-driven portraits animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1982–1991, 2023. 2

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

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

[34] Ivan Skorokhodov, Sergey Tulyakov, and Mohamed Elhoseiny. Stylegan-v: A continuous video generator with the price, image quality and perks of stylegan2, 2021. 6

[35] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 1, 3, 5, 6

[36] Michał Stypułkowski, Konstantinos Vougioukas, Sen He, Maciej Zieba, Stavros Petridis, and Maja Pantic. Diffused heads: Diffusion models beat gans on talking-face generation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5091–5100, 2024. 2

[37] Shuai Tan, Bin Ji, and Ye Pan. Emmn: Emotional motion memory network for audio-driven emotional talking face generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. 2

[38] Linrui Tian, Qi Wang, Bang Zhang, and Liefeng Bo. Emo: Emote portrait alive-generating expressive portrait videos with audio2video diffusion model under weak conditions. *arXiv preprint arXiv:2402.17485*, 2024. 2, 6, 4

[39] Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphaël Marinier, Marcin Michalski, and Sylvain Gelly. Fvd: A new metric for video generation. 2019. 6

[40] Cong Wang, Kuan Tian, Jun Zhang, Yonghang Guan, Feng Luo, Fei Shen, Zhiwei Jiang, Qing Gu, Xiao Han, and Wei Yang. V-express: Conditional dropout for progressive training of portrait video generation. *arXiv preprint arXiv:2406.02511*, 2024. 2

[41] Duomin Wang, Yu Deng, Zixin Yin, Heung-Yeung Shum, and Baoyuan Wang. Progressive disentangled representation learning for fine-grained controllable talking head synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 2, 6, 1

[42] Kaisiyuan Wang, Qianyi Wu, Linsen Song, Zhuoqian Yang, Wayne Wu, Chen Qian, Ran He, Yu Qiao, and Chen Change Loy. Mead: A large-scale audio-visual dataset for emotional talking-face generation. In *European Conference on Computer Vision*, pages 700–717. Springer, 2020. 6, 3

[43] Suzhen Wang, Lincheng Li, Yu Ding, Changjie Fan, and Xin Yu. Audio2head: Audio-driven one-shot talking-head generation with natural head motion. *arXiv preprint arXiv:2107.09293*, 2021. 2

[44] Huawei Wei, Zejun Yang, and Zhisheng Wang. Aniportrait: Audio-driven synthesis of photorealistic portrait animation. *arXiv preprint arXiv:2403.17694*, 2024. 2, 6, 8, 4

[45] Mingwang Xu, Hui Li, Qingkun Su, Hanlin Shang, Liwei Zhang, Ce Liu, Jingdong Wang, Luc Van Gool, Yao Yao, and Siyu Zhu. Hallo: Hierarchical audio-driven visual synthesis for portrait image animation. *arXiv preprint arXiv:2406.08801*, 2024. 2, 5, 6, 8, 7

[46] Sicheng Xu, Guojun Chen, Yu-Xiao Guo, Jiaolong Yang, Chong Li, Zhenyu Zang, Yizhong Zhang, Xin Tong, and Baining Guo. Vasa-1: Lifelike audio-driven talking faces generated in real time. *arXiv preprint arXiv:2404.10667*, 2024. 2

[47] Dongchao Yang, Songxiang Liu, Rongjie Huang, Chao Weng, and Helen Meng. Instructtts: Modelling expressive tts in discrete latent space with natural language style prompt. *TASLP*, 2024. 4

[48] Hani Yehia, Philip Rubin, and Eric Vatikiotis-Bateson. Quantitative association of vocal-tract and facial behavior. *Speech Communication*, 26(1-2):23–43, 1998. 4

[49] Wenxuan Zhang, Xiaodong Cun, Xuan Wang, Yong Zhang, Xi Shen, Yu Guo, Ying Shan, and Fei Wang. Sadtalker: Learning realistic 3d motion coefficients for stylized audio-driven single image talking face animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8652–8661, 2023. 2, 6, 8

[50] Yuang Zhang, Jiaxi Gu, Li-Wen Wang, Han Wang, Junqi Cheng, Yuefeng Zhu, and Fangyuan Zou. Mimicmotion: High-quality human motion video generation with confidence-aware pose guidance. *arXiv preprint arXiv:2406.19680*, 2024. 5

[51] Zhimeng Zhang, Lincheng Li, Yu Ding, and Changjie Fan. Flow-guided one-shot talking face generation with a high-resolution audio-visual dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3661–3670, 2021. 6, 3, 4

[52] Rui Zhen, Wenchao Song, Qiang He, Juan Cao, Lei Shi, and Jia Luo. Human-computer interaction system: A survey of talking-head generation. *Electronics*, 12(1):218, 2023. 1

[53] Daquan Zhou, Weimin Wang, Hanshu Yan, Weiwei Lv, Yizhe Zhu, and Jiashi Feng. Magicvideo: Efficient video generation with latent diffusion models. *arXiv preprint arXiv:2211.11018*, 2022. 2

[54] Xinfa Zhu, Yi Lei, Kun Song, Yongmao Zhang, Tao Li, and Lei Xie. Multi-speaker expressive speech synthesis via multiple factors decoupling. *arXiv preprint arXiv:2211.10568*, 2022. 4