ExpTalk: Diverse Emotional Expression via Adaptive Disentanglement and Refined Alignment for Speech-Driven 3D Facial Animation

Zhan Qu[1] , **Shengyu Zhang**[1*] , **Mengze Li**[1] , **Zhuo Chen**[2] ,
Chengfei Lv[3*] , **Zhou Zhao**[1*] and **Fei Wu**[1]
[1]Zhejiang University
[2]University of Electronic Science and Technology of China
[3]Alibaba Group
{quzhan, sy_zhang, mengzeli, zhaozhou, wufei}@zju.edu.cn, zhuochen@std.uestc.edu.cn,
chengfei.lcf@alibaba-inc.com

Abstract

Speech-driven 3D facial animation aims to create lifelike facial expressions that synchronize accurately with speech. Despite significant progress, many existing methods may focus on generating facial animation with a fixed emotional state, neglecting the diverse transformations of facial emotions under a given speech input. To solve this issue, we focus on exploring the refined alignment between speech representations and multiple domains in facial expression information. We aim to disentangle the spoken language and emotion facial priors from speech expressions, to guide the refinement of the facial vertices based on speech. To accomplish this objective, we propose ExpTalk, which first applies an Adaptive Disentanglement Variational Autoencoder (AD-VAE) to decouple facial expression aligned with spoken language and emotions of speech through contrastive learning. Then a Refined Alignment Diffusion (RAD) is employed to iteratively refine the decoupled facial expression priors through diffusion-based perturbations, producing facial animations that align with the emotional variations of the given speech. Extensive experiments prove the effectiveness of our ExpTalk by surpassing state-of-the-arts by a large margin.

1 Introduction

Speech-driven 3D facial animation aims to generate natural and realistic 3D facial animation highly aligned with the corresponding speech. With the rapid development of deep learning, existing speech-driven 3D facial animation methods efficiently generate relatively accurate facial expressions of spoken language through lip synchronization [Cudeiro *et al.*, 2019; Fan *et al.*, 2022; Peng *et al.*, 2023a; Xing *et al.*, 2023; Fan *et al.*, 2025]. Thanks to significant performance, these methods exhibit promising potential for application across a spectrum of fields, including digital avatars, virtual reality, interactive entertainment, and online meetings [Morishima, 1998; Tanaka *et al.*, 2022].

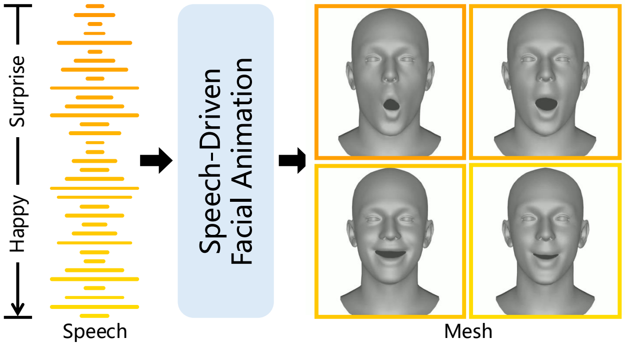

Figure 1: Given speech with emotional variations as input (left), we aim to generate facial animations with diverse emotional expressions (right), e.g., from surprise to happy.

Despite significant progress, many existing methods mainly focus on high-quality synchronization of facial language expressions, but overlook the fine-grained generation of facial emotional expressions. Such subtle facial emotions are another core component of facial expressions apart from facial spoken language [Murray and Arnott, 1993], playing a crucial role in daily interactions. Recognizing this, some methods [Daněček *et al.*, 2023; Peng *et al.*, 2023b] attempt to utilize a single emotional aspect in speech to aid in synthesizing the emotion in a complete facial animation. However, in the real world, human emotions are complex and diverse, which results in multiple emotions present in each complete speech needing to be reflected in the facial expression. For example, upon receiving an unexpected gift, individuals typically experience a moment of surprise quickly transitioning into happiness, a shift synchronously reflected in their speech and facial emotions, as shown in Figure 1. Rigid facial expressions aligned with the spoken language in speech but failing to fully match diverse speech emotions may potentially trigger the uncanny valley effect [Mori *et al.*, 2012].

To tackle the above issues, we propose a method for 3D facial animation generation, which decouples spoken language and emotions from the original facial expressions, followed by detailed refinements based on the corresponding speech information to produce new expected facial expressions. Our method has two goals: **(1) Facial Expression Decoupling.**

[*]Corresponding authors.

The facial expressions of spoken language and emotions are deeply fused. When individually refining them based on the corresponding information in speech, they tend to mutually interfere each other. Especially, when dealing with complex emotional transitions requiring detailed refinements, such adverse interference may be particularly significant. Thus, before detailed refinements, it is essential to decouple spoken language and emotions in facial expressions. **(2) Facial Expression Refinement.** When transferring facial expressions of spoken language and emotions, it is essential to make subtle refinements to numerous facial vertices one by one. This demand becomes more pronounced, particularly when swiftly transitioning between emotions. For this demand, we design deep damage to the facial features and deep reconstruction techniques to refine facial expressions.

Technically, we introduce the ExpTalk model for speech-driven 3D facial animation to achieve each objective, which consists of two target designs: **(1) Adaptive Disentanglement VAE (AD-VAE).** This module decouples the original facial expressions into two types of quantization codebooks, encapsulating spoken language and emotion priors. Leveraging extracted speech emotions, we adopt fine-grained contrastive learning to drive the training of AD-VAE for adaptive encodings disentangling. **(2) Refined Alignment Diffusion (RAD).** This module first utilizes the noise-injection encoder to make feature-dimension compression and iterative perturbations for facial priors extracted by AD-VAE. Later, RAD extracts speech features aligned with facial priors from multiple dimensions. With the refined speech features, we completely reconstruct facial priors to produce expected facial expressions synchronized with speech.

Our contributions are as follows:

- To the best of our knowledge, we take the early exploration of the emotion-shift problem in speech-driven 3D facial animation and propose ExpTalk to decouple the emotions from facial expressions, enabling fine-grained refinement of facial emotions based on speech emotion.
- We propose Adaptive Disentanglement VAE to decouple the facial features via contrastive learning, preventing mutual interference in the subsequent refinement of each facial prior. We propose Refined Alignment Diffusion to iteratively damage and reconstruct facial priors to align with corresponding speech.
- Extensive experiments prove the effectiveness of our ExpTalk model, demonstrating its potential and effectiveness in practical applications.

2 Related Work

2.1 Speech-Driven 3D Facial Animation

Recent work on speech-driven 3D facial animation methods [Cudeiro *et al.*, 2019; Peng *et al.*, 2023a] has achieved significant progress in generating highly synchronized lip animations based on speech. FaceFormer [Fan *et al.*, 2022] adopts a transformer-based [Vaswani, 2017] model to capture the relevant speech information, autoregressively generating continuous facial animations. CodeTalker [Xing *et al.*, 2023] leverages VQ-VAE to learn a discrete code space and employs a temporal autoregressive model to sequentially synthesize facial animations from input speech signals. UniTalker [Fan *et al.*, 2025] introduces a unified multi-head model to address annotation inconsistencies, enabling 3D facial animation across multi-domain datasets. However, these methods fail to adequately consider the characteristics of speech in conveying emotions, resulting in facial animations that still lack expressiveness.

To address this, recent studies [Kim *et al.*, 2024; Xie *et al.*, 2025] have incorporated emotional information as an additional condition to create emotional facial animations. EmoTalk [Peng *et al.*, 2023b] utilizes an emotion disentangling encoder with a cross-reconstruction loss to decouple emotion and content from speech, followed by an emotion-guided multi-head attention decoder to generate facial animations. EMOTE [Daněček *et al.*, 2023] employs a temporal variational autoencoder to learn motion priors and uses annotated emotional vector labels with a transformer encoder-decoder structures for animation generation. Despite these advances, these methods still struggle with refined alignment with real speech emotions.

2.2 Probabilistic Mapping of Diffusion Models

Recent advances in probabilistic modeling have enabled more effective cross-modal alignment in various domains [Li *et al.*, 2023a; Ji *et al.*, 2023; Li *et al.*, 2023b], among which diffusion models [Ho *et al.*, 2020; Song *et al.*, 2020] stand out due to their capability to model complex data distributions by iteratively adding noise and denoising. The extension to human-driven tasks [Tevet *et al.*, 2023; Zhu *et al.*, 2024] has motivated exploration in speech-driven 3D facial animation, where they outperform traditional methods in generating diverse expressions. FaceDiffuser [Stan *et al.*, 2023] first applied the diffusion framework to facial animation using audio features as conditions and employing a GRU decoder for animation generation. DiffSpeaker [Ma *et al.*, 2024] enhanced diversity of facial animations by integrating a transformer-based architecture with designed bias. Building on this, our work aims to leverage the iterative perturbation of diffusion models to globally capture high-dimensional facial feature distributions, enabling more refined cross-modal mapping.

3 Method

We aim to generate 3D facial animations synchronized with the given speech, which contains diverse emotions. The task can be formulated as follows: let $\boldsymbol{A}_{1:T} = (\boldsymbol{a}_1, \ldots, \boldsymbol{a}_T)$ denote the input speech sequence corresponding to T video frames, where each $\boldsymbol{a}_t \in \mathbb{R}^D$ contains D samples. Let $\boldsymbol{V}_{1:T} = (\boldsymbol{v}_1, \ldots, \boldsymbol{v}_T)$ denote the facial vertices sequence of length T, with each frame represented by V vertices, where $\boldsymbol{v}_t \in \mathbb{R}^{V \times 3}$. Our goal is to generate facial vertices sequence $\boldsymbol{V}_{1:T}$ with emotional variations based on speech $\boldsymbol{A}_{1:T}$.

As shown in Figure 2, ExpTalk follows a two-stage pipeline. *In the first stage*, we introduce Adaptive Disentanglement VAE (AD-VAE) to reconstruct the facial vertices sequence. Leveraging fine-grained contrastive learning with a pre-trained speech emotion feature extractor, this module adaptively disentangles the spoken language and emo-

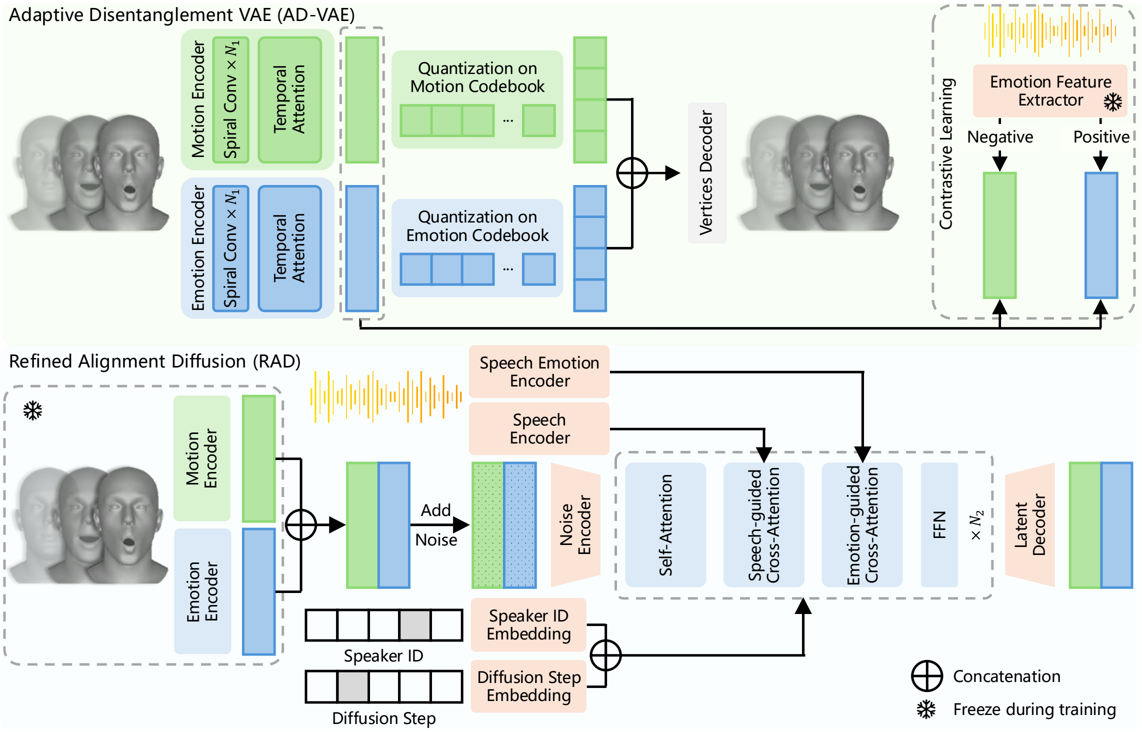

Figure 2: **Overview of ExpTalk.** In the first stage, AD-VAE employs contrastive learning to learn and disentangle the spoken language and emotion priors of facial vertices, quantized into two sub-codebooks. In the second stage, RAD extracts features from speech and recovers the noise-injected encodings through speech-guided and emotion-guided conditional attention mechanisms.

tional facial priors, which are then quantized into two sub-codebooks for the reconstruction of the facial vertices sequence. *In the second stage*, we propose Refined Alignment Diffusion (RAD), which refines facial expressions based on speech information by integrating speech-guided and emotion-guided conditional attention mechanisms to denoise the noise-injected encodings from encoders of AD-VAE.

During inference, the model ExpTalk takes the speech sequence $\boldsymbol{A}_{1:T}$ and speaker identity $\boldsymbol{p}$ as inputs. By iteratively denoising random noise, the model generates the facial animation $\hat{\boldsymbol{V}}_{1:T} = (\hat{\boldsymbol{v}}_1, \ldots, \hat{\boldsymbol{v}}_T)$. Formally,

$$\hat{\boldsymbol{V}}_{1:T} = \mathrm{ExpTalk}(\boldsymbol{A}_{1:T}, \boldsymbol{p}; \Theta), \tag{1}$$

where Θ denotes the learnable parameters of the model.

3.1 Adaptive Disentanglement VAE

In facial expressions, spoken language and emotions may be deeply fused, leading to mutual interference when separately refining them. The mutual interference may be more significant, especially when refining more diverse emotions. Inspired by [Zhang *et al.*, 2020; Wu *et al.*, 2024], Adaptive Disentanglement VAE (**AD-VAE**) is designed to disentangle facial expressions into two independent priors: spoken language and emotion. By embedding them into distinct quantization sub-codebooks, AD-VAE ensures their independence in the latent space. This disentanglement process is driven by fine-grained contrastive learning, which leverages speech emotion features as a supervisory signal to adaptively disentangle facial priors.

Model Design

AD-VAE consists of a motion encoder, an emotion encoder, and a facial vertices decoder. The motion and emotion encoders share the same structure and are designed to extract spoken language and emotional priors, respectively. Both encoders utilize spiral convolution operators to extract spatial features of facial vertices and temporal attention to capture the temporal dependencies within the vertices sequence. This design effectively integrates both the spatial correlations among vertices and the temporal dependencies of the sequence. The spiral convolution operator [Bouritsas *et al.*, 2019; Gong *et al.*, 2019] is a graph-based convolution that samples vertices in the spatial domain to capture local correlations. Given a convolution center v, the local neighborhood is described by $k-\mathrm{ring}(v)$, the set of vertices exactly k steps away from v, and $k-\mathrm{disk}(v)$, the union of all $i-\mathrm{ring}(v)$ for $i \leq k$. They are formally defined as:

$$\begin{aligned} 0-ring(v) &= v,\\ k-disk(v) &= \cup_{i=0,\ldots,k} i - ring(v),\\ (k+1)-ring(v) &= \mathcal{N}(k-ring(v)) \setminus k-disk(v), \end{aligned} \tag{2}$$

where $\mathcal{N}(\cdot)$ denotes the set of neighboring vertices of a given vertex. The spiral sequence is defined by concatenating several $k-ring(v)$ until reaching a predefined length. Using these sequences, the spiral convolution operator encodes local vertices information via a fully connected layer. Temporal attention is then applied to the encoded features, generating motion encoding $\boldsymbol{z}_{\mathrm{m}} \in \mathbb{R}^{T \times N \times C}$ and emotion encoding

$\boldsymbol{z}_\mathrm{e} \in \mathbb{R}^{T\times N\times C}$, collectively denoted as $\boldsymbol{z}_*$.

To ensure independence between the encodings, a pre-trained speech emotion feature extractor, emotion2vec [Ma *et al.*, 2023], is utilized to extract emotional features from the corresponding speech. During training, contrastive learning is applied to enhance the alignment between the facial emotion encoding and speech emotional features while reducing the similarity between the facial motion encoding and speech emotional features, achieving encoding disentanglement [Li *et al.*, 2024].

After extracting the encodings, the corresponding vectors $\mathbf{e}_i$ in the sub-codebook $\mathcal{E}_*$ are queried using a nearest-neighbor search operation for quantization:

$$\hat{\boldsymbol{z}}_* = \arg\min_{\mathbf{e}_i\in\mathcal{E}_*} \|\boldsymbol{z}_* - \mathbf{e}_i\|_2^2, \tag{3}$$

where $\|\cdot\|_2^2$ denotes the L^2 norm.

For decoding, the facial vertices decoder adopts a structure symmetric to the encoder. The quantized encodings are concatenated and mapped to their original dimensions, followed by temporal attention and spiral convolution operators to reconstruct the facial vertices sequence $\hat{\boldsymbol{V}}_{1:T}$.

Loss Functions

To fully train our AD-VAE, we introduce carefully designed loss functions, which consist of three parts: reconstruction loss $\mathcal{L}_\mathrm{rec}$, quantization loss $\mathcal{L}_\mathrm{qua}$, and contrastive loss $\mathcal{L}_\mathrm{con}$.

The reconstruction loss $\mathcal{L}_\mathrm{rec}$ ensures that the model can accurately reconstruct the sequence. This loss is computed as the L^1 loss between the ground truth facial vertices sequence and the predicted sequence:

$$\mathcal{L}_\mathrm{rec} = \frac{1}{T\times V}\sum_{t=1}^{T}\sum_{i=1}^{V}\left\|\boldsymbol{v}_{t,i} - \hat{\boldsymbol{v}}_{t,i}\right\|_1, \tag{4}$$

where T denotes the sequence length, and $\|\cdot\|_1$ denotes the L^1 norm used for distance calculation.

The quantization loss $\mathcal{L}_\mathrm{qua}$ is used to constrain the proximity between the unquantized latent encodings and the codebook vectors. This loss is defined as the weighted sum of codebook loss term and commitment loss term proposed from VQ-VAE [Van Den Oord *et al.*, 2017]:

$$\mathcal{L}_\mathrm{qua} = \|\hat{\boldsymbol{z}}_* - sg(\boldsymbol{z}_*)\|_2^2 + \beta\left\|\boldsymbol{z}_* - sg(\hat{\boldsymbol{z}}_*)\right\|_2^2, \tag{5}$$

where $sg(\cdot)$ denotes the gradient-stop operation.

The contrastive loss $\mathcal{L}_\mathrm{con}$ guides the disentanglement of motion and emotion encodings. To achieve this, a pre-trained speech emotion feature extractor is introduced to extract speech emotion features $\boldsymbol{e}\in\mathbb{R}^{T\times K}$, where K is the feature dimension. Emotion encodings of corresponding frame are paired with speech emotion features as positive samples, while motion encodings serve as negative samples. The loss is defined using InfoNCE [Oord *et al.*, 2018]:

$$\mathcal{L}_\mathrm{con} = -\frac{1}{T}\sum_{i=1}^{T}\log\frac{\exp(\mathrm{sim}(\boldsymbol{e}_i, \boldsymbol{z}_{\mathrm{e},i})/\tau)}{\sum_{j=1}^{N}\exp(\mathrm{sim}(\boldsymbol{e}_i, \boldsymbol{z}_{\mathrm{m},j})/\tau)}, \tag{6}$$

where τ denotes the temperature, and $\mathrm{sim}(\cdot,\cdot)$ denotes the cosine similarity, defined as $\mathrm{sim}(\boldsymbol{a},\boldsymbol{b}) = \frac{\boldsymbol{a}^\top\boldsymbol{b}}{\|\boldsymbol{a}\|\|\boldsymbol{b}\|}$.

The total loss for AD-VAE $\mathcal{L}_\mathrm{ADV}$ is given by:

$$\mathcal{L}_\mathrm{ADV} = \mathcal{L}_\mathrm{rec} + \lambda_\mathrm{qua}\mathcal{L}_\mathrm{qua} + \lambda_\mathrm{con}\mathcal{L}_\mathrm{con}, \tag{7}$$

where λ_qua and λ_con are hyperparameters.

3.2 Refined Alignment Diffusion

After decoupling spoken language and emotion priors of facial expressions, achieving seamless synchronization and expressiveness in facial animations requires precise cross-modal alignment with the corresponding speech information. To realize this objective, we design Refined Alignment Diffusion (**RAD**), a diffusion-based denoising network that iteratively refines spoken language and emotion priors of facial expressions to align with the corresponding speech information. We attempt to preserve high-frequency details and enable realistic, emotionally consistent facial animations synchronized with speech in the alignment process.

During training, the denoising network of RAD $f(\cdot)$ takes the noise-injected encodings $\boldsymbol{z}_\mathrm{m}^n$ and $\boldsymbol{z}_\mathrm{e}^n$, conditioned on the speech $\boldsymbol{A}_{1:T}$, speaker identity $\boldsymbol{p}$, and diffusion step n, and directly predicts the clean encodings $\hat{\boldsymbol{z}}_\mathrm{m}^0$ and $\hat{\boldsymbol{z}}_\mathrm{e}^0$,

$$\hat{\boldsymbol{z}}_\mathrm{m}^0, \hat{\boldsymbol{z}}_\mathrm{e}^0 = f(\boldsymbol{z}_\mathrm{m}^n, \boldsymbol{z}_\mathrm{e}^n, \boldsymbol{A}_{1:T}, \boldsymbol{p}, n). \tag{8}$$

In the inference phase, the high-noise version distributions of the motion and emotion encodings at diffusion step N, $p(\boldsymbol{z}_\mathrm{m}^N, \boldsymbol{z}_\mathrm{e}^N)$, are sequentially converted to the low-noise version distributions through a Markov chain, until the clean encoding distributions $p(\boldsymbol{z}_\mathrm{m}^0, \boldsymbol{z}_\mathrm{e}^0)$ are achieved:

$$\begin{aligned} p(\boldsymbol{z}_\mathrm{m}^0, \boldsymbol{z}_\mathrm{e}^0|\boldsymbol{A}_{1:T}, \boldsymbol{p}, N) &= p(\boldsymbol{z}_\mathrm{m}^N, \boldsymbol{z}_\mathrm{e}^N) \\ \prod_{n=1}^{N} p(\boldsymbol{z}_\mathrm{m}^{n-1}, \boldsymbol{z}_\mathrm{e}^{n-1}&|\boldsymbol{z}_\mathrm{m}^n, \boldsymbol{z}_\mathrm{e}^n, \boldsymbol{A}_{1:T}, \boldsymbol{p}, n), \end{aligned} \tag{9}$$

where $p(\boldsymbol{z}_\mathrm{m}^N)\sim\mathcal{N}(0,I)$ and $p(\boldsymbol{z}_\mathrm{e}^N)\sim\mathcal{N}(0,I)$ represent the initial distributions of motion and emotion encodings, modeled as standard normal distributions $\mathcal{N}(0,I)$ with a mean of 0 and an identity covariance matrix I. Subsequently, they are passed into the decoder of the trained AD-VAE to reconstruct the facial vertices sequence.

Model Design

RAD consists of encoders, conditional embeddings, and a backbone conditional transformer decoder. We use WavLM [Chen *et al.*, 2022] as the pre-trained speech encoder $E_a(\cdot)$ to extract speech features from $\boldsymbol{A}_{1:T}$ and emotion2vec [Ma *et al.*, 2023] as the pre-trained speech emotion encoder $E_e(\cdot)$ to extract emotion features. Both features are linearly transformed to match the dimensions required for the diffusion process. The noise input $\boldsymbol{z}^n = [\boldsymbol{z}_\mathrm{m}^n, \boldsymbol{z}_\mathrm{e}^n]$, which concatenates the motion and emotion noise-injected encodings, is processed by the noise encoder $E_n(\cdot)$, a linear layer that compresses high-dimensional noise to retain critical motion and emotion information.

The speaker identity embedding $E_p(\boldsymbol{p})$ and diffusion step embedding $E_t(n)$ are introduced as conditional embeddings into the model to enhance the model's capacity for capturing speaker style [Thambiraja *et al.*, 2023; Yu *et al.*, 2024] and diffusion step information.

Methods	3D-MEAD				VOCASET			
	MVE↓ (x10^{-4})	LVE↓ (x10^{-4})	EVE↓ (x10^{-5})	FDD↓ (x10^{-6})	MVE↓ (x10^{-5})	LVE↓ (x10^{-5})	EVE↓ (x10^{-6})	FDD↓ (x10^{-7})
FaceFormer [Fan *et al.*, 2022]	2.0250	6.8722	6.8028	3.8110	3.0653	7.5879	8.8491	3.8452
CodeTalker [Xing *et al.*, 2023]	1.9007	6.8714	6.7396	3.9093	3.1632	7.9097	9.0418	3.8115
FaceDiffuser [Stan *et al.*, 2023]	1.8882	6.1975	6.3530	3.9476	**1.7558**	4.5374	7.4725	4.5054
UniTalker [Fan *et al.*, 2025]	1.9628	5.7877	6.3427	3.5193	2.2663	5.1653	7.9655	4.7602
EMOTE [Daněček *et al.*, 2023]	1.7965	2.9881	7.7131	2.6517	3.6182	19.609	21.802	3.7362
EmoTalk [Peng *et al.*, 2023b]	1.7512	5.3816	7.0731	3.6462	2.3534	6.1929	9.7849	3.1588
DEEPTalk [Kim *et al.*, 2024]	1.5506	3.2658	7.0984	2.2827	5.7721	22.191	24.294	14.766
ExpTalk (Ours)	**1.4891**	**2.6947**	**4.4064**	**2.1790**	2.5932	**4.4744**	**5.4821**	**3.1522**

Table 1: Quantitative Evaluation Results on 3D-MEAD and VOCASET datasets. The best and the second best results are highlighted in bold and underlined.

Inspired by DiffSpeaker [Ma *et al.*, 2024], we adopt a conditional transformer decoder as the backbone for denoising. This decoder takes speaker identity $\boldsymbol{p}$ and diffusion step n as conditions, and incorporates speech-guided and emotion-guided conditional attention to progressively remove noise.

Specifically, the network consists of masked conditional self-attention, masked conditional cross-attention, and a feed-forward network. The query $\boldsymbol{Q}$ in the self-attention comes from the noise input encoding $E_n(\boldsymbol{z}^n) \in \mathbb{R}^{T \times N}$, while the keys $\boldsymbol{K}$ and values $\boldsymbol{V}$ are obtained by concatenating the speaker identity embedding $E_p(\boldsymbol{p})$, diffusion step embedding $E_t(n)$, and the noise input encoding $E_n(\boldsymbol{z}^n)$. Based on prior experience [Fan *et al.*, 2022; Ma *et al.*, 2024], we apply a symmetric temporal period mask $\mathcal{M}_s(i,j)$ to restrict attention to a small range within the current frame, while also considering the speaker identity and diffusion step information:

$$\mathcal{M}_s(i,j) = \begin{cases} 0, & 1 \leq j \leq 2, \\ \lfloor (i-j)/\mathbf{p} \rfloor, & 2 < j \leq i, \\ \lfloor (j-i)/\mathbf{p} \rfloor, & i < j \leq T+2, \end{cases} \tag{10}$$

where $\mathbf{p}$ denotes the temporal period, which corresponds to the frame rate. The operator $\lfloor \cdot \rfloor$ represents the floor function.

The speech-guided and emotion-guided conditional cross-attention mechanisms share an identical structure. Their keys $\boldsymbol{K}$ and values $\boldsymbol{V}$ are constructed by concatenating the speaker identity encoding $E_p(\boldsymbol{p})$, the diffusion step encoding $E_t(n)$, and either the speech features $E_a(\boldsymbol{A}_{1:T})$ or the emotion features $E_e(\boldsymbol{A}_{1:T})$. To ensure strong frame-level correlations between the speech and emotion features and their corresponding encodings, we apply an alignment mask $\mathcal{M}_c(i,j)$:

$$\mathcal{M}_c(i,j) = \begin{cases} 0, & \text{if } |i-j| \leq k \text{ or } j \in \{1,2\}, \\ -\infty, & \text{otherwise}, \end{cases} \tag{11}$$

where k denotes the frame-level alignment window size, controlling the range of temporal correlation between frames.

After multiple layers of the conditional transformer decoder, $\hat{\boldsymbol{z}}_{\mathrm{m}}^0$ and $\hat{\boldsymbol{z}}_{\mathrm{e}}^0$ are dimensionally recovered by the latent decoder $D_n(\cdot)$, which is implemented as a linear layer.

Loss Functions

The loss for RAD $\mathcal{L}_{\mathrm{RAD}}$ is defined to measure the difference between the predicted encodings after the denoising process and the ground truth encodings. We use the Huber loss to quantify this difference, as follows:

$$\mathcal{L}_{\mathrm{RAD}} = \mathbb{E}_{p(z_{\mathrm{m}}^n, z_{\mathrm{e}}^n)} \left[\left\| \hat{z}_{\mathrm{m}}^0 - z_{\mathrm{m}}^0 \right\|_H + \left\| \hat{z}_{\mathrm{e}}^0 - z_{\mathrm{e}}^0 \right\|_H \right], \tag{12}$$

where $\mathbb{E}_{p(z_{\mathrm{m}}^n, z_{\mathrm{e}}^n)}$ is the expectation over the noise encodings distribution, and $\| \cdot \|_H$ denotes the Huber loss [Huber, 1992].

4 Experiments

4.1 Experimental Settings

Datasets. In our experiments, we employ the 3D-MEAD dataset [Wang *et al.*, 2020; Daněček *et al.*, 2023], which provides high-quality facial expression data for training and evaluating emotion-driven facial animation methods. Each 3D facial mesh is registered to the FLAME topology [Li *et al.*, 2017], with 5023 vertices. The dataset includes speech and 3D reconstruction data from 46 English-speaking subjects across eight emotion categories: Neutral, Happy, Sad, Surprise, Fear, Disgust, Anger, and Contempt, with three intensity levels for each emotion except Neutral. Facial vertices sequences are sampled at 25 frames per second. To introduce emotional variations, we randomly concatenate the speech and facial vertices data. Due to the absence of official data configuration, we select 10 subjects for training and 3 unseen subjects for testing to ensure fairness.

To further evaluate our model, we employ the VOCASET dataset [Cudeiro *et al.*, 2019], which includes 480 speech samples and 3D facial reconstruction data from 12 subjects. The facial animations are originally sampled at 60 frames per second. We downsample them to 25 frames per second and randomly concatenate the speech and facial vertices data to align with the same configuration as the 3D-MEAD dataset. We follow the official data split for fair comparisons.

Implementation Details. For the training of ExpTalk, we use the Adam optimizer [Kingma, 2014] in both stages. AD-VAE is trained for 100 epochs with a batch size of 1 and a learning rate of 1×10^{-3}, taking approximately 10 hours. RAD is also trained for 100 epochs with a batch size of 1 and a learning rate of 2×10^{-5}, taking about 12 hours. During inference, we use 50 steps of the DDIM [Song *et al.*, 2020] sampler. Our framework is implemented with PyTorch and runs on an Nvidia RTX 4090 GPU.

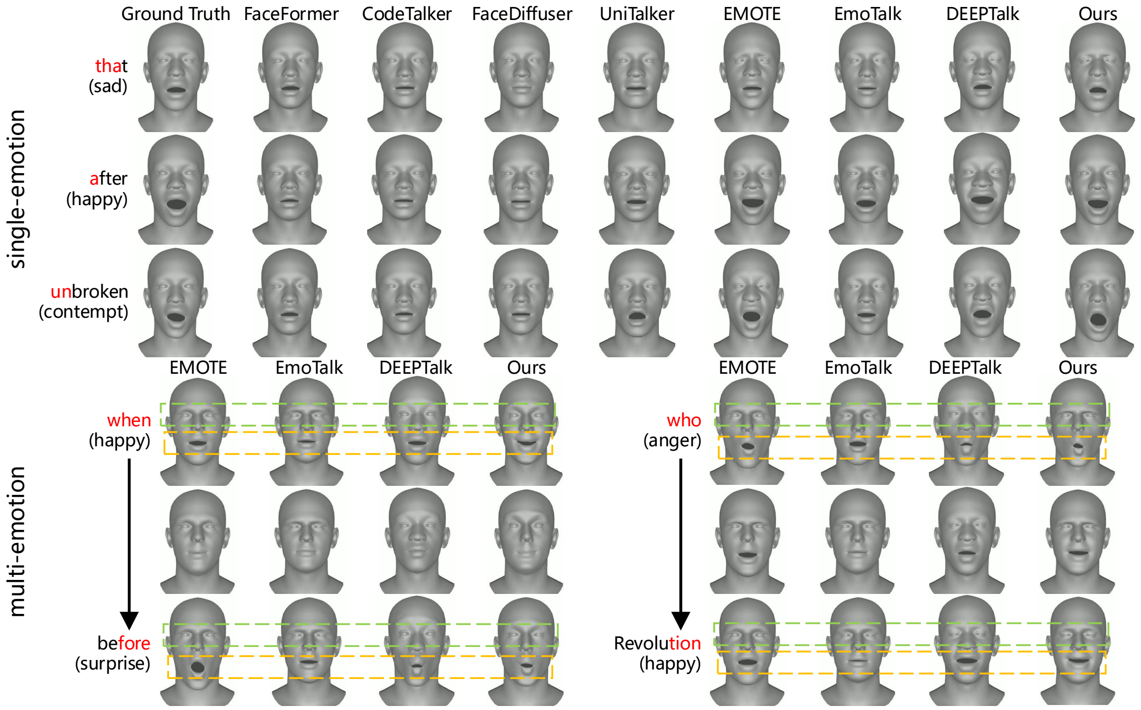

Figure 3: Qualitative evaluation results on single-emotion and multi-emotion speech. The green and orange boxes highlight the emotional variations on the upper face and lips, respectively.

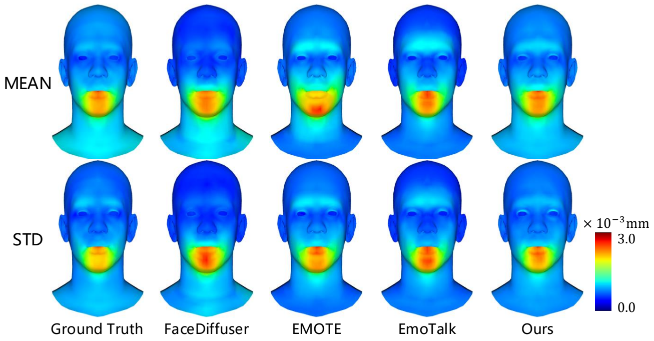

Figure 4: Heatmap visualization of mean and standard deviation: comparison of ground truth and animations from different methods.

Compared Baselines. We extend several state-of-the-arts as baselines for a comprehensive comparison: (1) non-emotion-driven methods: FaceFormer [Fan *et al.*, 2022], CodeTalker [Xing *et al.*, 2023], FaceDiffuser [Stan *et al.*, 2023], and UniTalker [Fan *et al.*, 2025]. (2) emotion-driven methods: EMOTE [Daněček *et al.*, 2023], EmoTalk [Peng *et al.*, 2023b], and DEEPTalk [Kim *et al.*, 2024].

4.2 Quantitative Evaluation

We follow the evaluation metrics used in CodeTalker [Xing *et al.*, 2023] and FaceDiffuser [Stan *et al.*, 2023], including Mean Vertex Error (MVE), Lip Vertex Error (LVE), and Upper-Face Dynamics Deviation (FDD). To evaluate the accuracy of facial emotion representation, we also adopt Emotional Vertex Error (EVE) from EmoTalk [Peng *et al.*, 2023b].

MVE measures overall accuracy via the Euclidean distance between generated and ground-truth facial vertex sequences. **LVE** measures accuracy in the lip region, reflecting lip synchronization, while **EVE** focuses on the upper face (e.g., forehead, eyes), capturing emotional expressiveness. **FDD** evaluates dynamic consistency by comparing the standard deviation of upper facial movement between generated and ground-truth sequences.

Table 1 presents the quantitative evaluation results. On the 3D-MEAD dataset, ExpTalk demonstrates outstanding performance across all metrics, showcasing its capability to generate emotionally expressive 3D facial animations. On the VOCASET dataset, ExpTalk leads in most metrics, although its MVE score is slightly lower than comparison models. This can be attributed to the characteristics of the VOCASET dataset, where speech samples are relatively short, and the dynamic range of vertex movements is limited. Such dataset distribution may place greater emphasis on static regions, which are less critical for evaluating dynamic emotional expressions. Despite this, ExpTalk excels in metrics such as LVE and EVE, which focus on dynamic facial regions, validating its ability to generate expressive and emotionally consistent facial animations aligned with speech.

4.3 Qualitative Evaluation

To further validate our method, we visualize and compare facial animations driven by single-emotion and multi-emotion speech. Single-emotion speech samples are from the 3D-MEAD dataset, while multi-emotion speech is synthesized by concatenating single-emotion speech clips with 1-second transitions. The results are shown in Figure 3.

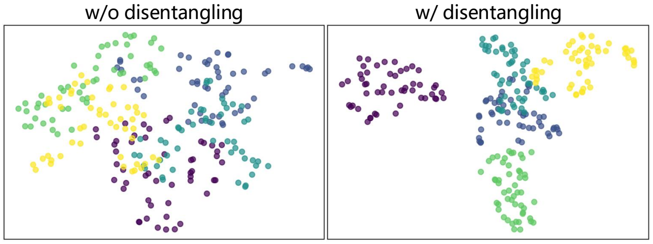

Figure 5: The t-SNE visualization of facial encodings under five emotions. The left figure shows encodings without disentanglement, while the right displays emotion encodings with disentanglement.

Methods	LipSync↑	Expressiveness↑	Realism↑
CodeTalker	2.6	2.0	2.4
UniTalker	3.5	2.1	2.9
EMOTE	3.5	3.6	3.3
EmoTalk	3.3	3.1	3.0
Ours	**3.6**	**3.8**	**3.7**

Table 2: User Study Results for Lip Synchronization, Emotional Expressiveness, and Visual Realism.

For single-emotion speech, non-emotion-driven methods leave the upper facial region relatively static and lack emotional expressiveness. Emotion-driven methods often generate overly exaggerated or indistinct expressions. In contrast, ExpTalk, through decoupling and refinement, generates facial animations with better lip synchronization and emotional expressiveness, closely matching the ground truth.

For multi-emotion speech, ExpTalk achieves emotional transitions in facial expressions, such as from happy to surprise or anger to happy. By aligning facial expressions with diverse speech emotions, our model produces more natural and expressive animations than other methods.

We also visualize heatmaps of the mean and standard deviation of facial animations generated by four methods compared to the ground truth. Figure 4 demonstrates that ExpTalk achieves superior performance, with upper facial movements more closely aligning with the ground truth. Additionally, We use t-SNE[Van der Maaten and Hinton, 2008] to visualize facial encodings with and without disentanglement, as shown in Figure 5. The results indicate that disentanglement enhances the distinction of different facial emotions, further validating its effectiveness in reconstructing facial expressions.

4.4 User Study

To evaluate ExpTalk's effectiveness, we conducted a user study comparing it with CodeTalker, UniTalker, EMOTE, and EmoTalk. Using 90 speech samples from the 3D-MEAD, RAVDESS [Livingstone and Russo, 2018], and online videos, covering diverse demographics, languages, and prosodic features. We generated facial animations driven by single and multiple emotional speech. Participants rated randomly presented videos on lip synchronization, emotional expressiveness, and visual realism using a 1–5 scale. The mean opinion scores are shown in Table 2. ExpTalk outperformed compared methods across all aspects, demonstrating its ability to generate expressive and emotionally rich facial animations.

Methods	MVE↓ $(x10^{-4})$	LVE↓ $(x10^{-4})$	EVE↓ $(x10^{-5})$
w/o conditional transformer decoder	1.9101	4.0953	7.4156
w/o emotion2vec	1.5407	2.7288	4.6375
w/o disentangle	1.6249	2.8212	4.5238
w/o diffusion	1.6043	2.8569	5.4564
Ours	**1.4891**	**2.6947**	**4.4064**

Table 3: Ablation Study.

4.5 Ablation Study

We conduct ablation studies on the 3D-MEAD dataset. The results are as shown in Table 3 and Figure 5.

Effect of the Key Components. We investigate the impact of the conditional Transformer decoder and emotion2vec. Replacing the conditional decoder with a standard Transformer decoder, we add speaker identity and diffusion step embeddings to the noise input and speech features. Results show that removing the conditional decoder significantly degrades multiple metrics, primarily because the model initially relies on audio information but gradually shifts focus to speaker identity as the noise is removed. Replacing emotion2vec with wav2vec [Baevski *et al.*, 2020] in both modules results in decreased emotional expressiveness, indicating that emotion2vec enhances fine-grained emotion modeling through its pretrained knowledge and effectively guides both AD-VAE disentanglement and RAD refinement.

Effect of the Model Architecture. We evaluate the necessity of facial feature disentanglement and the diffusion framework. Removing the emotion branch in AD-VAE leads to entanglement between emotional and non-emotional information in facial features, making it difficult for the model to clearly refine facial expressions and resulting in facial animations with ambiguous emotional characteristics. This observation is further validated by Figure 5. Replacing the diffusion framework with a Transformer decoder to predict facial priors causes a significant drop in EVE, demonstrating that the iterative denoising in the diffusion process plays a crucial role in the fine-grained alignment of facial priors.

5 Conclusion

In this work, we propose an innovative speech-driven 3D facial animation method, ExpTalk, aimed at generating more natural and emotionally expressive facial animations. The meticulously designed AD-VAE employs fine-grained contrastive learning to embed disentangled encodings into distinct quantization sub-codebooks. RAD then facilitates the refined alignment of cross-modal information between speech and facial priors, ensuring lip synchronization and emotional consistency in the generated speech-driven facial animations. Extensive qualitative and quantitative experiments validate the method's advantages on multiple benchmark datasets. We hope ExpTalk will inspire further research toward more realistic facial animation generation.

Ethical Statement

This research focuses on advancing speech-driven 3D facial animation technology, which has applications in fields such as virtual reality, digital avatars, and interactive entertainment. We commit to ensuring that the proposed methods are developed and applied responsibly, avoiding any misuse in areas such as deepfake creation or deceptive practices. Ethical considerations, including privacy and consent, are prioritized in all stages of data usage and model deployment.

Acknowledgements

This work is supported by National Natural Science Foundation of China (No. 62222211, U24A20326, 62402429), Key Research and Development Program of Zhejiang Province (No. 2024C03270), ZJU Kunpeng&Ascend Center of Excellence, Ningbo Yongjiang Talent Introduction Programme (2023A-397-G) and Young Elite Scientists Sponsorship Program by CAST (2024QNRC001).

References

[Baevski *et al.*, 2020] 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.

[Bouritsas *et al.*, 2019] Giorgos Bouritsas, Sergiy Bokhnyak, Stylianos Ploumpis, Michael Bronstein, and Stefanos Zafeiriou. Neural 3d morphable models: Spiral convolutional networks for 3d shape representation learning and generation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7213–7222, 2019.

[Chen *et al.*, 2022] Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518, 2022.

[Cudeiro *et al.*, 2019] Daniel Cudeiro, Timo Bolkart, Cassidy Laidlaw, Anurag Ranjan, and Michael J Black. Capture, learning, and synthesis of 3d speaking styles. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10101–10111, 2019.

[Daněček *et al.*, 2023] 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*, pages 1–13, 2023.

[Fan *et al.*, 2022] Yingruo Fan, Zhaojiang Lin, Jun Saito, Wenping Wang, and Taku Komura. Faceformer: Speech-driven 3d facial animation with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18770–18780, 2022.

[Fan *et al.*, 2025] Xiangyu Fan, Jiaqi Li, Zhiqian Lin, Weiye Xiao, and Lei Yang. Unitalker: Scaling up audio-driven 3d facial animation through a unified model. In *European Conference on Computer Vision*, pages 204–221. Springer, 2025.

[Gong *et al.*, 2019] Shunwang Gong, Lei Chen, Michael Bronstein, and Stefanos Zafeiriou. Spiralnet++: A fast and highly efficient mesh convolution operator. In *Proceedings of the IEEE/CVF international conference on computer vision workshops*, pages 0–0, 2019.

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

[Huber, 1992] Peter J Huber. Robust estimation of a location parameter. In *Breakthroughs in statistics: Methodology and distribution*, pages 492–518. Springer, 1992.

[Ji *et al.*, 2023] Wei Ji, Renjie Liang, Zhedong Zheng, Wenqiao Zhang, Shengyu Zhang, Juncheng Li, Mengze Li, and Tat-seng Chua. Are binary annotations sufficient? video moment retrieval via hierarchical uncertainty-based active learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 23013–23022, 2023.

[Kim *et al.*, 2024] Jisoo Kim, Jungbin Cho, Joonho Park, Soonmin Hwang, Da Eun Kim, Geon Kim, and Youngjae Yu. Deeptalk: Dynamic emotion embedding for probabilistic speech-driven 3d face animation. *arXiv preprint arXiv:2408.06010*, 2024.

[Kingma, 2014] Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[Li *et al.*, 2017] Tianye Li, Timo Bolkart, Michael J. Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4d scans. *ACM Trans. Graph.*, 36(6), November 2017.

[Li *et al.*, 2023a] Mengze Li, Han Wang, Wenqiao Zhang, Jiaxu Miao, Zhou Zhao, Shengyu Zhang, Wei Ji, and Fei Wu. Winner: Weakly-supervised hierarchical decomposition and alignment for spatio-temporal video grounding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 23090–23099, 2023.

[Li *et al.*, 2023b] Mengze Li, Haoyu Zhang, Juncheng Li, Zhou Zhao, Wenqiao Zhang, Shengyu Zhang, Shiliang Pu, Yueting Zhuang, and Fei Wu. Unsupervised domain adaptation for video object grounding with cascaded debiasing learning. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 3807–3816, 2023.

[Li *et al.*, 2024] Li Li, Wei Ji, Yiming Wu, Mengze Li, You Qin, Lina Wei, and Roger Zimmermann. Panoptic scene graph generation with semantics-prototype learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3145–3153, 2024.

[Livingstone and Russo, 2018] Steven R Livingstone and Frank A Russo. The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english. *PloS one*, 13(5):e0196391, 2018.

[Ma *et al.*, 2023] Ziyang Ma, Zhisheng Zheng, Jiaxin Ye, Jinchao Li, Zhifu Gao, Shiliang Zhang, and Xie Chen. emotion2vec: Self-supervised pre-training for speech emotion representation. *arXiv preprint arXiv:2312.15185*, 2023.

[Ma *et al.*, 2024] Zhiyuan Ma, Xiangyu Zhu, Guojun Qi, Chen Qian, Zhaoxiang Zhang, and Zhen Lei. Diffspeaker: Speech-driven 3d facial animation with diffusion transformer. *arXiv preprint arXiv:2402.05712*, 2024.

[Mori *et al.*, 2012] Masahiro Mori, Karl F MacDorman, and Norri Kageki. The uncanny valley [from the field]. *IEEE Robotics & automation magazine*, 19(2):98–100, 2012.

[Morishima, 1998] Shigeo Morishima. Real-time talking head driven by voice and its application to communication and entertainment. In *AVSP'98 International Conference on Auditory-Visual Speech Processing*, 1998.

[Murray and Arnott, 1993] Iain R Murray and John L Arnott. Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion. *The Journal of the Acoustical Society of America*, 93(2):1097–1108, 1993.

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

[Peng *et al.*, 2023a] Ziqiao Peng, Yihao Luo, Yue Shi, Hao Xu, Xiangyu Zhu, Hongyan Liu, Jun He, and Zhaoxin Fan. Selftalk: A self-supervised commutative training diagram to comprehend 3d talking faces. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 5292–5301, 2023.

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

[Song *et al.*, 2020] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

[Stan *et al.*, 2023] Stefan Stan, Kazi Injamamul Haque, and Zerrin Yumak. Facediffuser: Speech-driven 3d facial animation synthesis using diffusion. In *Proceedings of the 16th ACM SIGGRAPH Conference on Motion, Interaction and Games*, pages 1–11, 2023.

[Tanaka *et al.*, 2022] Hiroki Tanaka, Satoshi Nakamura, et al. The acceptability of virtual characters as social skills trainers: usability study. *JMIR human factors*, 9(1):e35358, 2022.

[Tevet *et al.*, 2023] Guy Tevet, Sigal Raab, Brian Gordon, Yoni Shafir, Daniel Cohen-or, and Amit Haim Bermano. Human motion diffusion model. In *The Eleventh International Conference on Learning Representations*, 2023.

[Thambiraja *et al.*, 2023] Balamurugan Thambiraja, Ikhsanul Habibie, Sadegh Aliakbarian, Darren Cosker, Christian Theobalt, and Justus Thies. Imitator: Personalized speech-driven 3d facial animation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 20621–20631, 2023.

[Van Den Oord *et al.*, 2017] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.

[Van der Maaten and Hinton, 2008] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[Vaswani, 2017] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.

[Wang *et al.*, 2020] 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.

[Wu *et al.*, 2024] Fei Wu, Tao Shen, Thomas Bäck, Jingyuan Chen, Gang Huang, Yaochu Jin, Kun Kuang, Mengze Li, Cewu Lu, Jiaxu Miao, et al. Knowledge-empowered, collaborative, and co-evolving ai models: The post-llm roadmap. *Engineering*, 2024.

[Xie *et al.*, 2025] Jiajian Xie, Shengyu Zhang, Mengze Li, Zhou Zhao, Fei Wu, et al. Ecoface: Audio-visual emotional co-disentanglement speech-driven 3d talking face generation. In *The Thirteenth International Conference on Learning Representations*, 2025.

[Xing *et al.*, 2023] Jinbo Xing, Menghan Xia, Yuechen Zhang, Xiaodong Cun, Jue Wang, and Tien-Tsin Wong. Codetalker: Speech-driven 3d facial animation with discrete motion prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12780–12790, 2023.

[Yu *et al.*, 2024] Hongyun Yu, Zhan Qu, Qihang Yu, Jianchuan Chen, Zhonghua Jiang, Zhiwen Chen, Shengyu Zhang, Jimin Xu, Fei Wu, Chengfei Lv, et al. Gaussiantalker: Speaker-specific talking head synthesis via 3d gaussian splatting. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 3548–3557, 2024.

[Zhang *et al.*, 2020] Shengyu Zhang, Tan Jiang, Tan Wang, Kun Kuang, Zhou Zhao, Jianke Zhu, Jin Yu, Hongxia Yang, and Fei Wu. Devlbert: Learning deconfounded visio-linguistic representations. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 4373–4382, 2020.

[Zhu *et al.*, 2024] Bingwen Zhu, Fanyi Wang, Tianyi Lu, Peng Liu, Jingwen Su, Jinxiu Liu, Yanhao Zhang, Zuxuan Wu, Guo-Jun Qi, and Yu-Gang Jiang. Zero-shot high-fidelity and pose-controllable character animation. In *IJCAI*, 2024.