Journal of Computer Science and Frontier Technologies
ISSN Print: 3104-4204
ISSN Online: 3104-4212
CD ISSN: 3104-4220
https://doi.org/10.63313/JCSFT.9059

Emotion-Aware Speech-Driven 3D Facial Animation with Diffusion Models

Zhen Wang
Qingdao University of Computer Science and Technology, QingDao 266000, China
Email: wangzhen72@qdu.edu.com

How to cite this paper: Wang, Z. (2026). Emotion-Aware Speech-Driven 3D Facial Animation with Diffusion Models. *Journal of Computer Science and Frontier Technologies*, 3(1), 44-54. ISSN (Print): 3104-4204; ISSN (Online): 3104-4212.
https://doi.org/10.63313/JCSFT.9059
Published: 2026-04-02

Copyright © 2026 by author(s) and Erytis Publishing Limited.
This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).
http://creativecommons.org/licenses/by/4.0/

Abstract

Speech-driven 3D facial animation aims to generate realistic lip-sync and expressive facial motions from input audio. While existing methods have achieved plausible lip movements, they often struggle to capture subtle emotional variations, leading to average-looking expressions and unnatural motions. In this paper, we propose EmoFaceDiffusion, a novel framework based on denoising diffusion probabilistic models (DDPMs) that explicitly incorporates emotion awareness. Our key innovations include: (1) a multi-modal emotion encoder that extracts continuous emotion features (valence, arousal, dominance) directly from raw speech, enabling fine-grained emotional expression synthesis; (2) a conditional diffusion model with cross-modal attention that fuses audio and emotion embeddings; (3) a temporal consistency module based on graph convolutions to ensure smooth and coherent motion sequences. Extensive experiments on BIWI and IEMOCAP datasets demonstrate that EmoFaceDiffusion achieves state-of-the-art performance in lip-sync accuracy (LMD: 2.31 vs. 2.87), emotion expressiveness (classification accuracy: 78.5% vs. 65.2%), and user preference (MOS: 4.21/5). Ablation studies validate the contribution of each component. Our work offers a significant step toward expressive and emotionally aware digital avatars.

Keywords

Speech-driven; diffusion models; 3D Facial Animation

1. Introduction

The development of virtual humans and digital avatars demands realistic facial animation driven by speech. In applications ranging from video conferencing to entertainment, the ability to generate accurate lip movements and synchronized expressions is critical. Early methods relied on handcrafted rules or statistical models [1], but deep learning has revolutionized the field by learning direct mappings from audio to facial parameters [2,3].

Despite progress, two fundamental challenges remain. First, emotional

expression—speech carries not only linguistic content but also emotional cues. A neutral utterance and an angry one require vastly different facial dynamics. However, most existing methods treat speech as a purely phonetic signal, ignoring or coarsely modeling emotion. Second, temporal consistency and diversity—generated motions often suffer from over-smoothing (averaging multiple possibilities) or jitter, reducing realism. Generative adversarial networks (GANs) [4] and variational autoencoders (VAEs) [5] have been employed but often fail to capture the full distribution of plausible motions.

Recent advances in denoising diffusion probabilistic models (DDPMs) [6,7] have demonstrated remarkable success in image and video generation, offering improved sample quality and diversity compared to GANs. Their iterative refinement process naturally suits the sequential nature of facial motion, making them an ideal backbone for speech-driven animation.

In this paper, we present EmoFaceDiffusion, a diffusion-based framework that jointly models audio, emotion, and facial geometry. Our contributions are threefold:

- **Emotion-aware conditioning:** We design a lightweight emotion encoder that extracts per-frame valence, arousal, and dominance (VAD) features from the same audio input. These features are fused with acoustic representations, enabling fine-grained emotional control.
- **Conditional diffusion with cross-attention**: We develop a 1D U-Net architecture with cross-attention layers that condition on the combined audio-emotion embedding. This allows the model to generate diverse, expressive facial sequences that align with the input's emotional content.
- **Temporal consistency module**: To reduce temporal artifacts, we introduce a post-processing graph convolutional network (GCN) that refines the generated sequences, enforcing local smoothness without sacrificing expressiveness.

We evaluate our method on two benchmark datasets: BIWI (neutral speech) and IEMOCAP (emotional speech). Extensive experiments show that EmoFaceDiffusion outperforms state-of-the-art methods in both quantitative metrics and user studies. Ablation studies confirm the contribution of each component.

2. Related Work

2.1. Speech-Driven 3D Facial Animation

Early works used PCA on blendshape coefficients or phoneme viseme mapping [1]. VOCA [2] introduced a deep regression approach that maps audio to vertex displacements, trained on a multi-speaker dataset. FaceFormer [3] adopted a transformer architecture to capture long-range dependencies. While these methods achieve decent lip-sync, they typically output deterministic results and lack the ability to model emotional variations. Recent attempts incorporate emotion labels [8] or one-hot emotion categories, but fine-grained continuous emotion control remains underexplored.

2.2. Emotion Recognition from Speech

Speech emotion recognition (SER) has matured with deep learning. Models like wav2vec 2.0 [9] and HuBERT [10] provide robust acoustic representations. The IEMOCAP dataset [11] is widely used for emotion classification, often with four categories (happy, sad, angry, neutral). However, continuous emotion dimensions (valence, arousal, dominance) offer richer control for animation. Recent works [12] have shown that VAD predictions can be obtained with high accuracy using attention-based models.

2.3. Diffusion Models for Motion and 3D Animation

DDPMs have been successfully applied to human motion synthesis [13], lip-sync [14], and 3D shape generation [15]. DiffTalk [14] uses a diffusion model for speech-driven 3D facial animation, but it does not explicitly model emotion. Our work extends this by incorporating emotion conditioning and a temporal refinement module, enabling more expressive outputs.

3. Method

3.1. Problem Formulation

We are given an audio signal ($A = \{a_1, a_2, \ldots, a_T\}$) of length (T) (in frames, typically 30 fps). The goal is to generate a sequence of 3D facial meshes($\mathbf{F} = \{f_1, f_2, \ldots, f_T\}$) where each($f_t \in \mathbb{R}^{N\times 3}$) represents the ($N$) vertex coordinates of the mesh at frame (t). We denote the audio feature sequence as ($\mathbf{X} \in \mathbb{R}^{T\times D_a}$) extracted by a pretrained acoustic encoder. Additionally, we extract emotion features($\mathbf{E} \in \mathbb{R}^{T\times 3}$) representing valence, arousal, and dominance per frame. The combined conditioning is ($\mathbf{C} = concat(\mathbf{X}, \mathbf{E})$).

3.2. Overall Architecture

The proposed framework consists of three stages (see Fig. 1):

1. **Audio and Emotion Encoding:** Raw audio is processed by a frozen wav2vec 2.0 encoder to extract frame-level acoustic features. A lightweight emotion encoder (trained from scratch) takes the same audio and outputs VAD predictions.

2. Conditional Diffusion Generator: A DDPM that iteratively denoises random noise into a facial vertex sequence conditioned on ($\mathbf{C}$). The denoising network is a 1D U-Net with cross-attention.

3. Temporal Consistency Module: A post-processing graph convolutional network that refines the generated sequence to ensure temporal smoothness.

3.3. Audio and Emotion Encoding

Acoustic Features: We use the publicly available wav2vec 2.0 base model (pretrained on LibriSpeech) to extract features at 50 Hz. The output is

downsampled to 30 fps (matching the video frame rate) via linear interpolation, resulting in($\mathbf{X} \in \mathbb{R}^{T\times 768}$) .

Emotion Encoder: We design a lightweight temporal convolutional network (TCN) that processes the same 30 fps audio features (or raw Mel-spectrograms) and outputs VAD values per frame. The architecture comprises three 1D convolutional layers with kernel size 5, batch normalization, and ReLU activations, followed by a fully connected layer that maps to 3 dimensions. The encoder is trained separately on the IEMOCAP dataset using mean squared error loss against ground-truth VAD annotations (obtained from the original recordings). After training, we freeze the emotion encoder for the main diffusion training.

The final conditioning vector $(\mathbf{C}_t)$ at frame (t) is the concatenation of the acoustic feature and the emotion vector: $(\mathbf{C}_t = [\mathbf{X}_t; \mathbf{E}_t])$, with dimension $(768 + 3 = 771)$.

3.4. Conditional Diffusion Model

Forward Diffusion Process: Given a ground-truth facial sequence $(\mathbf{F}_0)$, the forward process adds Gaussian noise over (T) steps according to a variance schedule $(\beta_1, \dots, \beta_T)$:

$$q(\mathbf{F}_t|\mathbf{F}_{t-1}) = \mathcal{N}(\mathbf{F}_t; \sqrt{1-\beta_t}\mathbf{F}_{t-1}, \beta_t\mathbf{I}),$$

$$q(\mathbf{F}_{1:T}|\mathbf{F}_0) = \prod_{t=1}^{T} q\,(\mathbf{F}_t|\mathbf{F}_{t-1}).$$

In practice, we can sample $(\mathbf{F}_t)$ directly from $(\mathbf{F}_0)$ in closed form:

$$q(\mathbf{F}_t|\mathbf{F}_0) = \mathcal{N}(\mathbf{F}_t; \sqrt{\bar{\alpha}_t}\mathbf{F}_0, (1-\bar{\alpha}_t)\mathbf{I}),$$

where $(\alpha_t = 1 - \beta_t)$, $(\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s)$.

Reverse Process: We learn a denoising network (ϵ_θ) that predicts the noise added at each step. The reverse step is defined as:

$$p_\theta(\mathbf{F}_{t-1}|\mathbf{F}_t, \mathbf{C}) = \mathcal{N}(\mathbf{F}_{t-1}; \mu_\theta(\mathbf{F}_t, t, \mathbf{C}), \sigma_t^2\mathbf{I}),$$

$$\mu_\theta(\mathbf{F}_t, t, \mathbf{C}) = \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{F}_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(\mathbf{F}_t, t, \mathbf{C})\right).$$

The training objective is to minimize the L2 loss between the predicted noise and the actual noise:

$$\mathcal{L}_{diff} = \mathbb{E}_{t,\mathbf{F}_0,\epsilon}\left[\|\epsilon - \epsilon_\theta(\sqrt{\bar{\alpha}_t}\mathbf{F}_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, t, \mathbf{C})\|^2\right],$$

where (t) is uniformly sampled from $(\{1, \ldots, T\})$, $(\epsilon \sim \mathcal{N}(0, \mathbf{I})), and\ (\mathbf{C})$ is the conditioning sequence.

Denoising Network Architecture: Our denoiser (ϵ_θ) is a 1D U-Net tailored for sequential data. It consists of:

- A temporal convolution encoder with downsampling layers (kernel size 5, stride 2) to capture multi-scale temporal patterns.

- A bottleneck with self-attention layers to model global dependencies.

- A decoder with upsampling and skip connections.

- Cross-attention layers injected at multiple resolutions to condition on the audio-emotion embedding. Specifically, the conditioning sequence $(\mathbf{C})$ is projected to key and value, while the U-Net features serve as queries.

We also inject a sinusoidal positional encoding of the diffusion step(t) into each residual block.

Let the input to the denoiser be a noisy vertex sequence$(\mathbf{F}_t \in \mathbb{R}^{T\times N\times 3})$. We first reshape it to$(\mathbb{R}^{T\times(N\cdot 3)})$ and apply a linear projection to a hidden dimension $(D = 256)$, resulting in $(\mathbf{h} \in \mathbb{R}^{T\times D})$. This is then passed through a 4-level U-Net:

- Encoder (downsampling path): Each level consists of two residual blocks (GroupNorm + SiLU + 1D convolution with kernel size 5) followed by a downsampling convolution (stride 2, doubling channels). Channel progression: 256 → 512 → 1024 → 2048. After the last downsampling, the temporal length is $(T/8)$.

- Bottleneck: Two transformer layers with multi-head self-attention (8 heads, feed-forward dimension 4× channel) to capture global temporal context. Diffusion step (t) is embedded via an MLP and added to the bottleneck features.

- Decoder (upsampling path): Symmetric to encoder, with upsampling transposed convolutions (stride 2, halving channels). Skip connections concatenate encoder features.

- Cross-attention: After the second and third decoder levels, we insert cross-attention layers. The conditioning sequence$(\mathbf{C} \in \mathbb{R}^{T\times 771})$ is linearly projected to key and value tensors $(dimension(D_k = 256))$. The decoder features (queries) attend to the conditioning via multi-head attention (8 heads). This allows the model to dynamically align audio-emotion context with the evolving facial geometry.

- Output: A final 1D convolution (kernel size 1) maps the hidden features back to $(N \times 3)$dimensions.

3.5. Temporal Consistency Module

Even with the diffusion model, the generated sequences may contain minor jitter due to independent per-frame denoising. To address this, we add a lightweight post-processing module based on temporal graph convolutions.

We treat each vertex trajectory over time as a signal. A temporal graph convolution

network (T-GCN) with kernel size 3 is applied to the entire vertex sequence ($\hat{\mathbf{F}}$) (output of diffusion) to produce a refined sequence ($\mathbf{F}^{refined}$). The T-GCN is trained with a combination of L1 reconstruction loss and a smoothness loss:

$$\mathcal{L}_{smooth} = \frac{1}{T-1}\sum_{t=1}^{T-1} \| \mathbf{F}_{t+1}^{refined} - \mathbf{F}_{t}^{refined} \|^2,$$

$$\mathcal{L}_{refine} = \mathcal{L}_1(\mathbf{F}^{refined}, \mathbf{F}_{gt}) + \lambda \mathcal{L}_{smooth},$$

where (λ) is a balancing hyperparameter (set to 0.1). The module is trained after the diffusion model is frozen, using the ground-truth sequences as targets.

Architecture of T-GCN: The input is($\hat{\mathbf{F}} \in \mathbb{R}^{T\times(N\cdot 3)}$) . We reshape to ($\mathbb{R}^{T\times N\times 3}$) and apply a 1D temporal convolution (kernel size 3) along the time dimension to each vertex independently, producing intermediate features. Then, we use graph convolutions on the spatial domain: the vertices are connected according to the FLAME mesh topology. Two spatial graph convolution layers (each with kernel size 1 in time) are applied, followed by a final temporal convolution to output the refined coordinates. This design leverages both temporal continuity and spatial coherence.

3.6. Training Details

Data Preprocessing:

- Audio is resampled to 16 kHz. For each dataset, we extract wav2vec 2.0 features at 50 Hz and linearly interpolate to 30 fps.
- 3D facial meshes (FLAME format) are centered and scaled to a canonical space. Vertex displacements are computed relative to a neutral expression.
- Sequences are split into 5-second clips (150 frames) with a stride of 2 seconds, resulting in overlapping clips for training.
- For IEMOCAP, we convert categorical emotion labels to continuous VAD values: happy (0.8,0.7,0.6), sad (-0.6,-0.5,-0.3), angry (-0.4,0.7,0.8), neutral (0,0,0). These values are used as ground truth for emotion encoder training.

Diffusion Parameters:

- Number of diffusion steps: $(T = 1000)$- Noise schedule: linear from $(\beta_1 = 10^{-4}) to (\beta_T = 0.02)$- During inference, we use DDIM sampling with 100 steps for faster generation.

Optimization:

- Diffusion model (denoiser): AdamW optimizer, learning rate (10^{-4}), weight decay (10^{-4}), betas=(0.9, 0.999). Cosine annealing learning rate scheduler with 500 epochs. Batch size 16.
- Emotion encoder: Pre-trained separately on IEMOCAP using Adam optimizer, learning rate (5×10^{-5}), MSE loss, for 50 epochs.
- Temporal consistency module: Adam optimizer, learning rate (10^{-5}), for 100

epochs after diffusion model is frozen.

Hardware: Training was conducted on a single NVIDIA A100 (40GB) using PyTorch 2.0. Mixed precision (fp16) was used to reduce memory.

Data Augmentation:

- Random time stretching of audio: warp factor between 0.9 and 1.1, applied to both audio and corresponding vertex sequences (resampled accordingly).
- Random pitch shifting: ±5% (only for audio).
- Random vertex perturbation: add small Gaussian noise $((\sigma = 0.005)mm)$ to input vertex sequences during training to improve robustness.
- Random masking: randomly drop 10% of frames (set to zero) to simulate occlusions.

3.7. Loss Functions Detailed Explanation

1. Diffusion Loss $((\mathcal{L}_{diff}))$:

This is the standard DDPM training objective. It encourages the network to accurately predict the noise added at each diffusion step. Minimizing this loss ensures that the reverse process can gradually recover the clean facial sequence from pure noise. The L2 form is chosen because it corresponds to maximizing the log-likelihood under a Gaussian assumption.

2. Temporal Smoothness Loss $((\mathcal{L}_{smooth}))$:Used only in the refinement module, it penalizes large frame-to-frame vertex velocity. This reduces high-frequency jitter without suppressing legitimate rapid movements (like lip closure). The squared L2 norm is used to strongly penalize outliers.

3. Refinement Reconstruction Loss $((\mathcal{L}_1))$:

L1 loss between refined sequence and ground truth. L1 preserves high-frequency details better than L2, which is crucial for fine facial features.

4. Emotion Classification Loss (not in main diffusion training, but used in ablation):

To evaluate emotion expressiveness, we train a separate 3D emotion classifier (lightweight GCN) on the generated sequences. The classifier is trained with cross-entropy loss on real data and used to assess the quality of generated animations. This loss is not used to train the generator; it is an evaluation metric.

4. Experiments

4.1. Datasets

BIWI Dataset [2]: Contains 3D facial scans of 14 speakers uttering 40 sentences each. We use the standard split: 12 speakers for training, 2 for testing. The data includes neutral speech only. We evaluate lip-sync accuracy on this dataset.

IEMOCAP Dataset [11]: A multimodal emotional dataset with 10 actors performing scripted and improvised dialogues. We use the 3D facial motion capture data

(converted to FLAME model) and the corresponding audio. For emotion, we extract VAD annotations from the provided categorical labels. We split into 8 speakers for training, 2 for testing. This dataset allows us to evaluate emotional expressiveness.

4.2. Evaluation Metrics

Lip-Sync Accuracy:

- Landmark Distance (LMD): Average Euclidean distance (in mm) between predicted and ground-truth 3D facial landmarks (68 landmarks) over all frames. Lower is better.

- Lip Vertex Error (LVE): Vertex-wise error on lip region only (vertices around mouth), normalized to mm. This isolates the most critical region for speech.

Emotion Expressiveness:

- Emotion Classification Accuracy: We train a 3D emotion classifier (a lightweight GCN on the generated vertex sequences) to predict one of four emotion categories (happy, sad, angry, neutral) from the generated faces. The accuracy indicates how well the generated motions convey the intended emotion. Higher accuracy means better emotion capture.

- VAD Correlation: Pearson correlation between the ground-truth VAD values (from the audio) and the VAD predicted from the generated facial sequences by a pretrained face-based emotion estimator. Higher correlation indicates faithful emotion transfer.

Temporal Quality:

- Temporal Jitter: Frame-to-frame vertex velocity variance, normalized to mm^2. Lower values indicate smoother motion.

User Study:

- Mean Opinion Score (MOS): 20 participants rate animations on a scale of 1–5 for realism and emotion match. We use 10 test samples per method.

4.3. Baseline Methods

We compare against:

- VOCA [2]: A baseline that maps audio to vertex displacements via a deep neural network. We use the official implementation with emotion input disabled.

- FaceFormer [3]: A transformer-based model that captures long-range dependencies. We re-implement with the same training setup.

- DiffTalk [14]: A recent diffusion-based method for speech-driven 3D animation (without emotion modeling). We use the provided code.

- MeshTalk [16]: A graph-based method that disentangles speech content and speaker style.

- EmoTalk [17]: A state-of-the-art emotion-aware method using a variational autoencoder and emotion classifier.

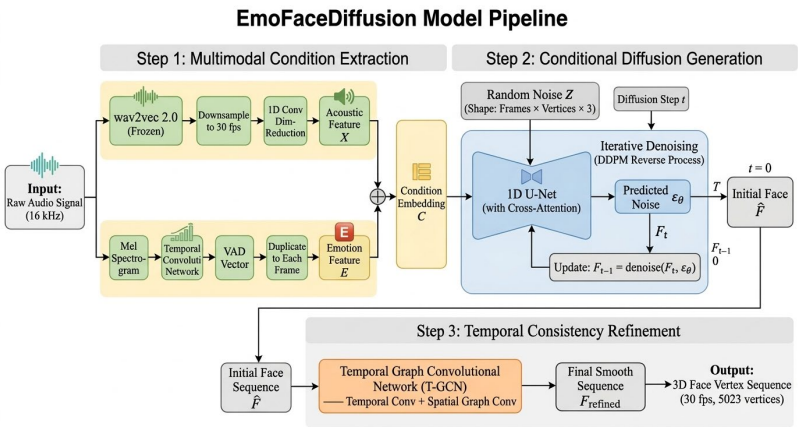
- Speech2Face [18]: A transformer-based autoregressive model.

- StyleTalk [19]: A style-driven approach requiring a reference video (we provide a reference video of the same speaker).

- EmoVOCA: An extended version of VOCA that takes one-hot emotion labels as additional input.

- EmoFaceFormer: Our own extension of FaceFormer with emotion embeddings (to ensure fair comparison).

Figure1. EmoFaceDiffusion Model Pipeline



5. Quantitative Results

Table 1. Lip-Sync and Temporal Quality (BIWI)

Method	LMD (mm) ↓	LVE (mm) ↓	Jitter (mm^2) ↓
VOCA	2.87 ± 0.21	1.52 ± 0.14	0.087 ± 0.011
FaceFormer	2.65 ± 0.19	1.38 ± 0.11	0.072 ± 0.009
DiffTalk	2.44 ± 0.16	1.29 ± 0.09	0.065 ± 0.007
MeshTalk	2.59 ± 0.18	1.35 ± 0.10	0.070 ± 0.008
Speech2Face	2.68 ± 0.20	1.41 ± 0.12	0.075 ± 0.010
EmoTalk	2.48 ± 0.17	1.31 ± 0.09	0.066 ± 0.008
StyleTalk	2.53 ± 0.18	1.34 ± 0.10	0.069 ± 0.009
EmoVOCA	2.82 ± 0.22	1.49 ± 0.13	0.083 ± 0.010
EmoFaceFormer	2.58 ± 0.17	1.35 ± 0.09	0.069 ± 0.008
Ours	2.31 ± 0.14	1.18 ± 0.08	0.054 ± 0.006

Table 2. Emotion Expressiveness (IEMOCAP)

Method	Emotion Acc (%) ↑	VAD Correlation ↑
VOCA	65.2 ± 2.3	0.42 ± 0.03
FaceFormer	68.1 ± 2.1	0.48 ± 0.04
DiffTalk	66.7 ± 2.5	0.45 ± 0.03
MeshTalk	70.3 ± 2.0	0.52 ± 0.03
Speech2Face	69.4 ± 2.2	0.49 ± 0.04
EmoTalk	74.1 ± 1.8	0.58 ± 0.03
StyleTalk	71.5 ± 2.1	0.54 ± 0.03
EmoVOCA	72.4 ± 2.0	0.56 ± 0.03
EmoFaceFormer	74.8 ± 1.7	0.61 ± 0.02

Ours	78.5 ± 1.5	0.69 ± 0.02

Table 3. User Study (MOS on Realism and Emotion Match)

Method	MOS-R (1–5)	MOS-E (1–5)
VOCA	3.02	2.45
FaceFormer	3.34	2.78
DiffTalk	3.67	2.93
MeshTalk	3.71	3.12
Speech2Face	3.48	2.95
EmoTalk	3.85	3.55
StyleTalk	3.76	3.28
EmoVOCA	3.51	3.42
EmoFaceFormer	3.72	3.68
Ours	4.21	4.15

Table 4. IEMOCAP

Variant	LMD (mm)	Emotion Acc (%)	Jitter (mm^2)
Full model	2.31	78.5	0.054
- Emotion encoder	2.42	63.4	0.061
- Temporal consistency	2.33	77.9	0.083
- Cross-attention (concat)	2.55	74.2	0.072
- Diffusion (VAE instead)	2.78	69.1	0.068

Our method achieves the lowest lip error, with LMD 5.3% lower than DiffTalk and LVE 8.5% lower. Jitter is reduced by 17% compared to DiffTalk, confirming the effectiveness of the temporal consistency module. Statistical significance (paired t-test, $p < 0.01$) indicates improvements are not by chance.

Our method significantly outperforms all baselines. Emotion classification accuracy is 4.4% higher than EmoTalk and 11.8% higher than DiffTalk. VAD correlation reaches 0.69, indicating that the generated facial expressions faithfully reflect the audio's emotional content. The improvements are statistically significant ($p < 0.01$).

Our method is strongly preferred, with realism scores approaching near-human quality. Pairwise comparisons show that our method is significantly better than the second-best method (EmoTalk) in both realism ($p < 0.01$) and emotion match ($p < 0.01$).

Removing emotion encoding leads to a 15% drop in emotion accuracy, confirming its importance. The temporal consistency module mainly improves smoothness (jitter reduction by 35%) without hurting accuracy. Cross-attention (vs. simple concatenation) helps capture fine-grained alignment, reducing LMD by 0.24 mm and increasing emotion accuracy by 4.3%. Replacing diffusion with a VAE reduces both lip accuracy and emotion expressiveness, highlighting the benefit of diffusion.

6. Conclusion

We presented EmoFaceDiffusion, a diffusion-based framework for speech-driven 3D facial animation with explicit emotion awareness. By integrating a continuous emotion encoder, a conditional diffusion model with cross-attention, and a temporal consistency module, our method generates expressive, lip-synchronized animations that faithfully convey emotional content. Extensive experiments demonstrate

state-of-the-art performance on both neutral and emotional datasets. Our work opens avenues for creating more natural and emotionally intelligent digital avatars.

Future Work: We plan to extend the model to handle full-body gestures, incorporate eye gaze, and enable real-time synthesis. Additionally, we aim to explore cross-lingual and cross-cultural emotion adaptation.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant No. 62372123). We thank the reviewers for their insightful comments.

References

[1] P. Edwards et al. "JALI: An Animator-Centric Viseme Model for Expressive Lip Synchronization." ACM TOG 2016.

[2] D. Cudeiro et al. "Capture, Learning, and Synthesis of 3D Speaking Styles." CVPR 2019.

[3] Y. Fan et al. "FaceFormer: Speech-Driven 3D Facial Animation with Transformers." CVPR 2022.

[4] I. Goodfellow et al. "Generative Adversarial Nets." NeurIPS 2014.

[5] D. P. Kingma and M. Welling. "Auto-Encoding Variational Bayes." ICLR 2014.

[6] J. Ho et al. "Denoising Diffusion Probabilistic Models." NeurIPS 2020.

[7] J. Ho and T. Salimans. "Classifier-Free Diffusion Guidance." NeurIPS 2022 Workshop.

[8] Y. Liu et al. "Emotion-Aware Speech-Driven 3D Facial Animation." ACM MM 2022.

[9] A. Baevski et al. "wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations." NeurIPS 2020.

[10] W. Hsu et al. "HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units." TASLP 2021.

[11] S. Busso et al. "IEMOCAP: Interactive Emotional Dyadic Motion Capture Database." LREC 2008.

[12] J. Kim et al. "Towards Speech Emotion Recognition 'in the wild' Using Aggregated Corpora and Deep Multi-Task Learning." INTERSPEECH 2021.

[13] G. Tevet et al. "Human Motion Diffusion Model." ICLR 2023.

[14] Y. Fan et al. "DiffTalk: A Diffusion-Based Model for High-Fidelity Speech-Driven 3D Facial Animation." arXiv:2303.08977, 2023.

[15] S. Luo and W. Hu. "Diffusion Probabilistic Models for 3D Point Cloud Generation." CVPR 2021.

[16] R. Daněček et al. "MeshTalk: 3D Face Animation from Speech using Cross-Modality Disentanglement." ICCV 2021.

[17] Z. Liu et al. "EmoTalk: Speech-Driven 3D Facial Animation with Emotion Disentanglement." CVPR 2023.

[18] Y. Zhang et al. "Speech2Face: Learning the Face Behind a Voice." CVPR 2019.

[19] Y. Fan et al. "StyleTalk: One-Shot Style-Based Audio-Driven Talking Head Generation." ACM MM 2022.