Speech-Driven 3D Facial Animation with Natural Head Movements

KUIYUAN SUN, Institute of Information Science, Beijing Key Laboratory of Advanced Information Science and Network Technology, Beijing Jiaotong University, Beijing, China
JICHAO ZHANG, School of Computer Science, Ocean University of China, Qingdao, China
WEI WANG and YAO ZHAO, Institute of Information Science, Beijing Key Laboratory of Advanced Information Science and Network Technology, Beijing Jiaotong University, Beijing, China
NICU SEBE, Department of Information and Computer Science, University of Trento, Trento, Italy

Speech-driven 3D facial animation has been studied for a long time, yet many challenges still prevent achieving truly natural results. One major issue is that current methods often overlook head motion during speech. To address this, we conducted a preliminary investigation and identified several key obstacles: the scarcity of 3D facial animation datasets with head motion and the limitations of regular sequence prediction models (e.g., Transformer, LSTM, and GRU), which are designed for discrete dynamic sequences and do not align well with continuous 3D head motion sequences. To solve these issues, we propose a head motion prediction module. This module uses audio and the initial motion state of the mesh to predict head motion. By employing ordinary differential equations (ODE) to model continuous dynamic sequences, it predicts head movements that closely resemble real head motion, making the animation more realistic and natural. Additionally, recognizing that the head motion generation given audio is not a one-to-one mapping problem, we introduce a noise module to help the head motion prediction module generate varied head motions given the same audio input. We also observed that previous facial animation methods primarily focus on generating vertices for the mouth region but use a single model to generate the entire face. This approach wastes some of the model's fitting capacity on other regions. To solve this problem, we propose a cascaded mesh generation module that uses two modules to separately generate the vertex of mouth region and other facial regions. Extensive experiments and a perceptual user study show that our approach outperforms existing methods and produces relatively natural head motion.

CCS Concepts: • **Applied computing → Performing arts**; • **Human-centered computing → Visualization techniques**; • **Information systems → Data management systems**;

This work was supported in part by the National Key R&D Program of China (No. 2021ZD0112100), National NSF of China (No. 62120106009), National NSF of China (No. 62372033), Fundamental Research Funds for the Central Universities (No. 2022XKRC015), China Academy of Traditional Chinese Medicine Science and Technology Innovation Project (CAM-KJHT-2024-0026), and Shandong Provincial Natural Science Foundation—Excellent Youth Program (Overseas Track) (No. 2026HWYQ-029).

Authors' Contact Information: Kuiyuan Sun, Institute of Information Science, Beijing Key Laboratory of Advanced Information Science and Network Technology, Beijing Jiaotong University, Beijing, China; e-mail: 20112001@bjtu.edu.cn; Jichao Zhang, School of Computer Science, Ocean University of China, Qingdao, China; e-mail: zhang163220@gmail.com; Wei Wang, Institute of Information Science, Beijing Key Laboratory of Advanced Information Science and Network Technology, Beijing Jiaotong University, Beijing, China; e-mail: weiwang2@bjtu.edu.cn; Yao Zhao (corresponding author), Institute of Information Science, Beijing Key Laboratory of Advanced Information Science and Network Technology, Beijing Jiaotong University, Beijing, China; e-mail: yzhao@bjtu.edu.cn; Nicu Sebe, Department of Information and Computer Science, University of Trento, Trento, Italy; e-mail: sebe@disi.unitn.it.

ACM 1551-6865/2026/7-ART207
https://doi.org/10.1145/3811818

Additional Key Words and Phrases: 3D Facial Animation, Sequence Prediction, 3D Mesh Generation

ACM Reference format:
Kuiyuan Sun, Jichao Zhang, Wei Wang, Yao Zhao, and Nicu Sebe. 2026. Speech-Driven 3D Facial Animation with Natural Head Movements. *ACM Trans. Multimedia Comput. Commun. Appl.* 22, 7, Article 207 (July 2026), 22 pages.
https://doi.org/10.1145/3811818

1 Introduction

Three-dimensional facial animation has garnered significant attention over the years owing to its diverse applications in virtual reality, film production, and gaming industries. The intricate connection between speech patterns and facial expressions, especially the nuanced movements of the lips, allows for sophisticated facial animations that accurately reflect the nuances of spoken language. This capability not only enhances realism in digital characters but also enriches user experiences across various interactive media platforms.

The early works mainly make attempts to build mapping rules between phonemes and their visual counterpart using different methods, including the autoregressive method [13], categorical latent space [40], and discrete primitives [56]. Despite the relentless efforts of researchers and the significant advancements in recent speech-driven facial animation techniques, we are still far from generating realistic human-like talking videos.

In real speaking scenarios, the head typically exhibits rhythmic, subtle movements. These head movements are crucial for generating human-like talking videos. Unfortunately, current 3D speaking mesh datasets [10, 38] lack head motion information, thus recent animation methods [10, 13, 40, 56] focus on facial movements, particularly lip motions, and neglect the head movement. To tackle these limitations, we use a 3D face reconstruction method [14] to extract head poses from 2D datasets. Generating head motion can be seen as a sequence prediction based on a given audio and initial head mesh motion state. We observed that the sequence prediction baselines (transformer-based models, **Long Short-Term Memory (LSTM)**-based models, and **Gated Recurrent Unit (GRU)**-based models) [36, 54, 66] tend to predict unstable, mechanical head movements. This is because head motion is essentially a continuous process, and current methods lack adequate modeling in this aspect, often generating dynamic sequences without considering continuity.

In this article, we propose a speech-driven facial animation model that generates facial animation videos with realistic head movements. As illustrated in Figure 1, this method initially generates a talking mesh with a static head driven by a segment of speech. Subsequently, it controls the head movement of this talking mesh. Our model consists of two main modules: **Head Motion Prediction Model (HMPM)** to generate head motion and **Mesh Generation Model (MGM)** to generate a mesh with synchronized lip shape (Figure 2). First, our HMPM uses neural **Ordinary Differential Equations (ODE)** over traditional sequence models to model continuous dynamic sequences for head motion prediction. Additionally, since the relationship between audio and head motion during speech is not a one-to-one mapping where the same words can be spoken with different head movements, we also propose a **Noise Disturbance Model (NDM)** to help our HMPM to generate diverse head motion sequences. Second, our MGM is designed for better lip and audio synchronization. We propose a cascaded talking mesh generation module to fully utilize the fitting capacity, consisting of dual encoders, each focusing on different regions of the face. The mouth region mesh generation module uses audio features to generate a coarse mesh with synchronized lip shape. This output is then fed into the next mesh generation module, which refines the coarse mesh based on audio features and mesh features from the previous module,

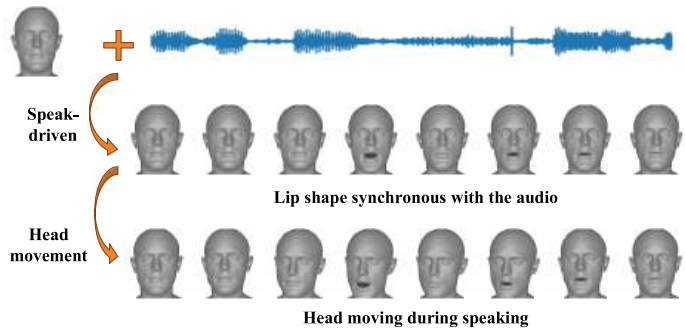

Fig. 1. Audio signal is used to generate speech-driven facial animation, and the corresponding head motion sequence is also generated. More experimental results videos can be found in the supplementary materials.

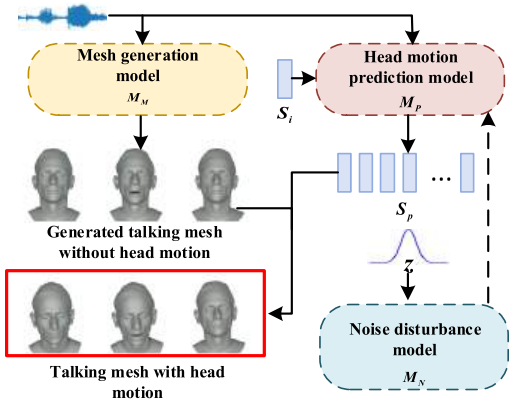

Fig. 2. Overview of the proposed model. It consists of an MGM M_M, an HMPM M_P, and an NDM M_N. M_M inputs audio and generates a talking mesh sequence. M_P takes an initial head motion state S_i and audio as inputs and outputs a prediction sequence S_p. M_N inputs noise samples z to control the diverse generation of M_P. (The dashed line indicates optional input.)

making the generated mesh more harmonious and natural. This design allows each module to focus on its sub-task, maximizing their fitting ability and achieving more accurate audio and lip synchronization.

Our work offers several key contributions:

—We propose a speech-driven facial animation model that generates videos with high lip synchronization and natural head movements.
—We propose utilizing ODEs to model continuous dynamic sequences, along with a NDM to enhance the diversity in generating head motions.
—We introduce a cascaded mesh generation module that uses dual encoders to separately generate the vertex of mouth region and other facial regions.
—Experiments demonstrate that our method achieves state-of-the-art results in both lip synchronization and natural head motion.

2 Related Works

2.1 Speech-Driven Facial Animation

With advancements in image generation techniques [20, 41, 55, 60], 3D reconstruction methods [12, 14, 29, 59], face detection methods [27, 65], and the release of extensive facial datasets [7], the field of speech-driven facial animation [10, 13, 40, 46, 56] has grown significantly.

Speech-driven facial animation generally falls into two categories: 2D and 3D animations. The study of 2D facial animation began earlier and still attracts many researchers. Traditional 2D facial animation [43, 63, 64] is usually based on generative adversarial networks [16] or variational autoencoders [21]. Due to the lack of high-quality datasets and less advanced image and video generation technology at that time, these methods faced issues like low-resolution outputs and insufficient image detail. However, with the development of generation technology [49] and the availability of high-quality datasets [1, 7, 8, 35], 2D facial animation methods have improved, now capable of producing high-quality and high-resolution videos.

Despite these advancements, 2D facial animation is limited to single-view video generation. To address this limitation, some researchers have turned to 3D-based methods, including mesh [18, 30, 50], **Neural Radiance Fields (NeRF)** [33], 3D Gaussian splatting [25], and signed distance fields (SDF) [2, 15, 22]. NeRF- [17, 61] and SDF-based methods [15] require multi-view videos or images of specific individuals and can only generate videos of those specific identities. Moreover, NeRF methods are time-consuming. In contrast, mesh-based methods have been widely studied due to their faster generation speed and flexibility, which are not restricted to specific individuals. These methods can be rendered in 2D or combined with other video generation technologies [39, 57, 58] to create multi-view talking head videos.

Different from 2D methods [32, 51, 64] and other 3D methods (NeRF- and SDF-based methods), mesh-based speech-driven facial animation methods rely on datasets that are easier to obtain compared to traditional 2D talking head methods. The datasets for these methods can be downloaded from video web sites and preprocessed [43]. In contrast, mesh-based speech-driven facial animation requires high-quality scanned 3D mesh datasets. Fortunately, several relevant datasets are available to support research on speech-driven talking meshes [10, 26, 38, 46].

Mesh-based speech-driven facial animation methods have evolved through several stages. In its early stages, VOCA [10] introduced a dataset of scanned talking facial meshes and a facial animation model, proposing a regressive model using a simple multi-layer perceptron structure. Considering facial animation as a regression task is inherently ill-posed. Therefore, there is a significant room for improvement in the results generated by VOCA. To solve this problem, the following methods treat facial animation as a generation task and better performance has been obtained. FaceFormer [13], using a transformer model, excels in handling sequence features, achieving superior lip-audio synchronization performance. CodeTalker [56] addresses the challenge of generating facial animations with neutral expressions by constructing a discrete latent code space to represent facial mesh expressions and introducing a new loss function to disentangle features from different modalities, preventing inefficient use of audio information. FaceDiffuser [44] attempts to use the diffusion model to generate the talking mesh. It uses audio as conditional information and generates the whole talking mesh through the denoise procedure.

However, the facial animations generated by current mesh-based methods are still unnatural and lack realism compared to real talking videos. Many factors contribute to this, such as the lack of emotional change and head motion. The Imitator [47] tries to solve this problem by learning a personalized speaking style. However, research on head movement during speech is still lacking. In this work, we focus on this problem and propose a facial animation method. It achieves

state-of-the-art performance on lip-sync and can generate head motion sequences using audio and the initial head motion state through a sequence prediction model.

2.2 Time Sequence Prediction

Time sequence prediction has long been a significant area of academic research, playing a crucial role in applications such as climate modeling [34], biological sciences and medicine [48], and commercial decision-making in retail and finance [3]. It is also used in predicting body joint movements [24, 37, 53].

Traditional methods for time series analysis relied on expert-guided models, such as autoregression and exponential smoothing. These approaches required significant domain-specific expertise and predefined parameters. However, the advent of modern machine learning methods has changed this landscape by enabling models to automatically identify patterns from data, thereby reducing the need for expert knowledge and decreasing dependency on predefined parameters.

This evolution has been further advanced by the development of sequence processing models like **Recurrent Neural Network (RNN)**, LSTM, and Transformers. Coupled with increased data availability and computing power, these deep learning methods have become essential for time series forecasting.

Early deep learning methods primarily employed convolutional neural networks and RNNs [36, 52, 62]. These approaches generally involved processing short sequences and making iterative predictions to generate the desired output. Recently, transformer-based models [28, 54, 66] have gained prominence due to their ability to effectively handle contextual information and capture sequence-related features, demonstrating superior performance in this domain. The advancement of transformers for time sequence prediction primarily focuses on two areas: improving the attention mechanism and enhancing decoder design.

Regarding the attention mechanism, the main challenge is to ensure its effectiveness. Variants of effective attention mechanisms can be broadly categorized into two types. The first type is Temporal Sparse Attention, which reduces complexity by using predefined patterns to sparsify attention [9]. Although this approach decreases computational demands, it is constrained by its fixed structures and may not fully capture all relevant correlations. The second type is Frequency Domain Attention, which integrates decomposition blocks with methods like the Fast Fourier Transform or other frequency analysis techniques to identify connections across sequences [6, 67].

For the decoder design, initially, autoregressive encoders were used for time sequence prediction [23]. Although these encoders performed well in natural language processing tasks, they resulted in unsatisfactory performance for time sequence prediction due to error accumulation, as mathematically proven in [45]. To address this, Informer [66] introduced non-autoregressive decoding to avoid error accumulation and improve efficiency. This approach was also adopted in subsequent models such as Autoformer [54], FEDformer [67], and Scaleformer [42].

Though the time series prediction models have great development, they are not suitable for head motion sequence prediction task. Unlike conventional long-term sequence prediction tasks, head motion sequence prediction emphasizes both accuracy and stability. The length of the head motion sequence to be predicted is often much longer than the initial head motion sequence, necessitating multiple predictions to achieve the desired length. Existing prediction methods tend to generate sequences that deviate from the natural head motion patterns as the number of predictions increases. To solve this problem, we develop a model to generate head motion sequences that control the natural movement of talking meshes. Generating a head motion sequence (3D head pose parameters) can be viewed as predicting head motion using audio input. To enhance model stability, we use a neural ODE to model a continuous dynamic sequence and propose an HMPM.

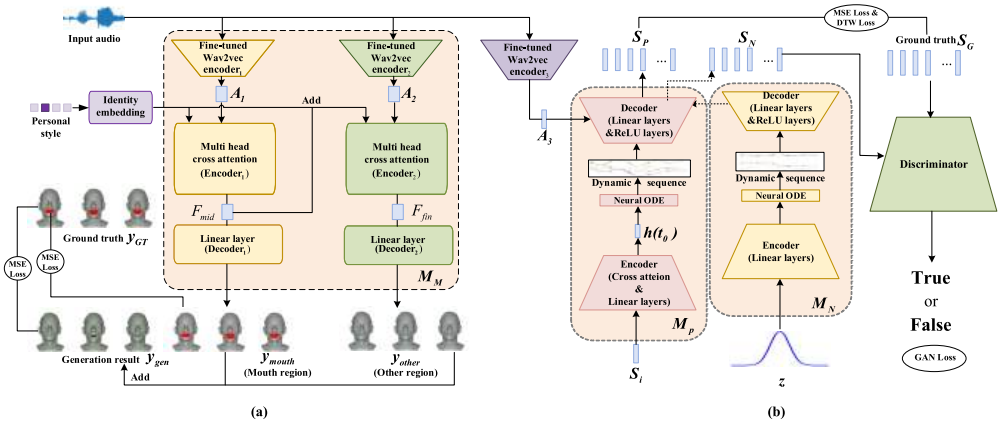

Fig. 3. (a) Training of the MGM M_M. (b) Training of the Head Motion Sequence Prediction Model M_P and NDM M_N. When training the NDM M_N, a multi-scale discriminator is adopted to perform adversarial training.

3 Method

Figure 2 illustrates the overall architecture, comprising a mesh generative model M_M designed to generate a talking mesh without head motion. Additionally, an HMPM M_P augments the system by predicting head motions associated with speech. To introduce diversity in head motion generation, an NDM M_N is integrated into M_P. Below, we provide detailed descriptions of each model.

3.1 MGM

The detailed architecture of the MGM M_N is shown in Figure 3(a). Given the diversity in speaking styles among individuals, we adopt a method where each person's unique style is represented by a one-hot vector. These vectors are then mapped to corresponding identity embeddings. The wav2vec encoder$_1$ takes audio feature A_1 and identity embedding as inputs and generates a mesh y_{mouth} with a mouth region similar to the Ground Truth. Then, the wav2vec encoder$_2$ uses the audio feature A_2 and output of the Encoder$_1$ to generate a mesh y_{other}, where the regions outside the mouth are similar to the Ground Truth. The outputs of two encoders are added to produce the final result. Finally, the y_{mouth} and y_{other} are fused to get the generation result y_{gen}.

Two loss functions are adopted to train the MGM M_M. The first loss function, mouth loss L_M, is used to train the Encoder$_1$ to generate the mouth region. It is defined as follows:

$$\mathcal{L}_M = \|mask_M(y_{GT}) - mask_M(y_{mouth})\|_2, \tag{1}$$

where $mask_M$ denotes the mask indicating the mouth region of the mesh, and y_{GT} represents the Ground Truth mesh.

The second loss function ensures the model generates meshes with all regions similar to the Ground Truth mesh. It is defined as follows:

$$\mathcal{L}_{all} = \left\|y_{GT} - y_{gen}\right\|_2. \tag{2}$$

Although L_M enables Encoder$_1$ to generate a mesh with accurate lip synchronization, other regions may not be well formed. Encoder$_2$ produces a mesh to compensate for the limitations of Encoder$_1$, and L_{all} ensures that the entire model generates an accurate overall mesh.

The final loss used to train M_M is shown in following equation:

$$\mathcal{L} = \lambda_1 \mathcal{L}_M + \lambda_2 \mathcal{L}_{all}. \tag{3}$$

3.2 HMPM

The architecture of the HMPM M_P is illustrated in Figure 3(b). It comprises an encoder, a decoder, and a neural ODE module [5], which are used to model continuous dynamic sequences.

First, the encoder processes the initial head motion state S_i to generate an initial value $h(t_0)$ for the neural ODE. Then, the neural ODE produces a dynamic sequence. Additionally, the wav2vec encoder$_3$ processes the input audio to obtain the audio feature A_3. Finally, the decoder uses the dynamic sequence and this audio feature A_3 to predict a head motion sequence S_P. The decoder can also use noise generated by the noise distribution module M_N to create diverse head motion sequences S_N.

Why choose neural ODE over traditional sequence prediction models? In previous models, the dynamic sequences generated by the encoder often lack continuity, resulting in discrete representations. However, head motion during speech is a continuous process. This discretization introduces discrepancies between the actual head motion sequence and the predicted outcomes, thereby impacting the accuracy and smoothness of head motion prediction and generation. To tackle this issue, it is crucial to model the dynamic sequence as continuous. Neural ODEs are preferred for modeling continuous dynamic sequences because they leverage the power of differential equations to provide a more natural and accurate representation of time-evolving phenomena, such as head motion during speech, compared to traditional sequence prediction models that rely on discrete time intervals. Therefore, we employ neural ODE to simulate a continuous dynamic sequence.

In our implementation, we adopt a fixed-step explicit Euler method to solve the ODE, as shown in Equation (4). This choice prioritizes training stability and computational efficiency and aligns well with our uniformly sampled motion data. While we acknowledge that this discretization shares mathematical similarities with a residual network, the core distinction lies in the "continuous-first" modeling perspective: we define a continuous dynamical system (via g) whose behavior is then approximated numerically. This perspective emphasizes learning smooth dynamics. Although adaptive solvers (e.g., DOPRI5) offer advantages in precision for certain tasks, the fixed-step Euler scheme suffices for our goal of capturing robust dynamic trends in head motion.

To briefly introduce the principle of neural ODE, we start by assuming the existence of a continuous function $H(t)$. Neural ODE aims to find a sequence $h(t_i)$ that closely approximates this function, given an initial value $h(t_0)$ and a gradient model g parameterized by a neural network. This is accomplished through the following equations:

$$\begin{aligned} h\left(t_{i+1}\right) &= h\left(t_i\right) + g\left(h\left(t_i\right)\right) \times \left(t_{i+1} - t_i\right), \\ g\left(h\left(t_i\right)\right) &= \frac{dH\left(t_i\right)}{dt}, \\ h\left(t_0\right) &= H\left(t_0\right), \end{aligned} \tag{4}$$

where t_i represents some discrete points on the t coordinate axis. The model g takes $h(t_i)$ as input and outputs derivative of $H(t_i)$.

The neural ODE takes the final item in the output feature sequence of the encoder as $h(t_0) \in R^{1\times 3}$ and solves a dynamic sequence $h(t_i) \in R^{n\times 3}$ according to a time sequence t_i where $i \in 0, 1, 2, \ldots, n$. n is the length of the time sequence t. Although the resulting $h(t_i)$ sequence is obtained through discrete steps, it is derived from a continuous dynamic model. This approach produces sequences that are more continuous in representation than those from fully discrete models (e.g., LSTM and Transformer), thus allowing for the generation of sequences that more closely resemble natural head motion.

Two loss functions are utilized to train the sequence prediction model: MSE loss L_{MSE} and **Dynamic Time Warping (DTW)** loss L_{DTW}. L_{MSE} guides the model to generate results S_P that

closely resemble the Ground Truth S_G numerically and is defined as:

$$\mathcal{L}_{MSE} = \|S_G - S_P\|_2. \tag{5}$$

The DTW loss [11] is employed to align the predicted sequences with the Ground Truth. It calculates the similarity between two sequences and utilizes it to guide the model. Unlike L_{MSE}, which focuses on making the values of the predicted sequence similar to the Ground Truth, L_{DTW} guides the model to learn the changing trend of the Ground Truth.

The final loss used to train M_P is shown in following equation:

$$\mathcal{L} = \lambda_3 \mathcal{L}_{MSE} + \lambda_4 \mathcal{L}_{DTW}. \tag{6}$$

3.3 NDM

As shown in Figure 3(b), the NDM M_N comprises an encoder and a decoder. It takes in Gaussian random noise sample z and outputs a noise list to induce variations in the motion prediction model M_P for generating different motion sequences. Similar to the HMPM, M_N also integrates a neural ODE between the encoder and decoder. The neural ODE utilizes the output of the encoder as the initial value $h(t_0)$ and predicts a noise dynamic sequence $h(t)$. The decoder shares a similar structure with the decoder of M_P but has one less LN-ReLU layer. The decoder generates a list containing the output of every layer in the decoder. Each layer's output is subsequently added to the corresponding layers in the decoder of M_P, excluding the final layer.

As there is no specific Ground Truth available to guide the training, we employ a multi-scale discriminator to perform the adversarial training. Throughout the training procedure, the parameters of M_P remain fixed, with the discriminator and M_N actively engaging in gradient backpropagation. To mitigate the risk of the discriminator dominating the training process and potentially hindering effective learning of M_N, discriminator updates its gradient only once every two training iterations. The training loss $\mathcal{L}_{adv}$ is defined as follows:

$$\begin{aligned}\mathcal{L}_{adv} = &\mathbf{E}_{z_i \sim P_{data}(Z)}[log(1-D(M_P(S_i, A_3, M_N(z_i))))] \\ &+ \mathbf{E}_{S_G \sim P_{data(S_G)}}[log(D(S_G))],\end{aligned} \tag{7}$$

where z_i denotes the input noise sample, and S_G represents the Ground Truth motion sequence.

A regularization term is also introduced to prevent the generated sequences from becoming overly similar and thereby losing diversity. It is defined as:

$$\begin{aligned}\mathcal{L}_N &= -\mathcal{L}_{DTW}(S_{N_1}, S_{N_2}), \\ S_{N_1} &= M_P(S_i, A_3, M_N(z_1)), \\ S_{N_2} &= M_P(S_i, A_3, M_N(z_2)),\end{aligned} \tag{8}$$

where S_{N_1} and S_{N_1} are two noise samples generated by M_N. This strategy aims to discourage excessive similarity in the evolution trends of the two sequences.

The final loss used to train M_N is shown in the following equation:

$$\mathcal{L} = \lambda_5 \mathcal{L}_{adv} + \lambda_6 \mathcal{L}_N. \tag{9}$$

4 Experiment

4.1 Experimental Settings

Implementation Details. Following previous works [4, 10, 13, 40, 46, 56], audio is sampled at 16 kHz and aligned with each frame using a centered 5-frame window. For head motion generation, we predict the future sequence $S_P \in \mathbb{R}^{50\times3}$ from an initial state $S_i \in \mathbb{R}^{n\times3}$ (with $n = 10$ in comparisons, though $n = 1$ suffices) and the corresponding audio chunk of size $1 \times 3{,}200$ (25 fps video).

Training Details. The MGM is trained with Adam for 100 epochs (batch size 1, learning rate 1×10^{-4}). The head motion predictor and NDM are trained sequentially with Adam for 200 epochs (batch size 10, lr 1×10^{-4}): first the predictor and then frozen to train the noise model. Hyperparameters: $\lambda_1 = 1.0$, $\lambda_2 = 0.5$, $\lambda_3 = 0.8$, $\lambda_4 = 0.1$, $\lambda_5 = 0.1$, $\lambda_6 = 0.5$. All experiments run on a single NVIDIA 3090 GPU. For head motion baselines, we additionally use DTW loss during training.

Datasets. Mesh generation is evaluated on VOCASET [10] and Multiface [19] (BIWI is unavailable). VOCASET contains 480 sequences (12 subjects, 60 fps, 3–4 s, 5,023 vertices) with official splits. Multiface provides 13 subjects, each speaking 50 sentences, with 7,306 vertices (cropped to 6,172, identity-based splits, normalized to $[-1, 1]$). For head motion learning, we use 2D video datasets RAVDESS [31] (7,356 segments, 24 actors, varied emotions) and VoxCeleb [7] (22,496 videos, more diverse motion). Head motion sequences are extracted via FOMM [43] (cropping+alignment) and DECA [14] (3D pose), then normalized.

Baselines. For mesh generation, we compare with VOCA [10], FaceDiffuser [44], FaceFormer [13], CodeTalker [56], DiffPoseTalk [46], and Perceptual 3D Talk [4]. Identity-conditioned methods use all training IDs for unseen subjects. For head motion, we compare sequence models: Informer [66], Autoformer [54], and QuaterNet (LSTM/GRU) [36]. To ensure fairness, all use our decoder and are trained with DTW loss. Qualitative evaluation with head motion is performed on VOCA only (due to DECA compatibility with Multiface).

Metrics. Lip-sync quality is measured by **Lip Vertex Error (LVE)**, the maximum L2 error across all lip vertices. Overall face quality is assessed by **Other Vertex Error (OVE)**, the maximal L2 error on non-lip vertices. **Upper-Face Dynamics Deviation (Upper-FDD)** [56] quantifies temporal variation:

$$\mathrm{FDD}(m_{1:T}, \hat{m}_{1:T}) = \frac{1}{|V_U|} \sum_{v \in V_U} \left(\mathrm{std}(\|m^v_{1:T}\|_2) - \mathrm{std}(\|\hat{m}^v_{1:T}\|_2) \right), \tag{10}$$

where V_U indexes upper-face vertices and $\mathrm{std}(\cdot)$ computes temporal standard deviation. For head motion, we report MSE, MAE, and DTW loss. Note that LVE magnitudes may differ from those in DiffPoseTalk [46] due to implementation details; we ensure consistent evaluation across all compared methods.

4.2 Quantitative Evaluation

Speech-Driven Mesh Generation. The results are shown in Table 1. We can observe that our method performs better than other methods in lip-sync error, other facial region error, and Upper-FDD. This indicates that our method can maintain lip and audio synchronization while keeping high consistency between the predicted upper-face expressions together with the trend of facial dynamics and those of the Ground Truth. On the VOCASET dataset [7], CodeTalker [13] performs second only to ours in all aspects, due to its use of discrete primitives to model the facial motion space, which enhances the realism of motion synthesis. The performance of DiffPoseTalk [46] and Perceptual 3D Talk [4] is also notably strong on this dataset, with both methods achieving results that are secondary and very close to our method's, and with Perceptual 3D Talk showing a slight edge among the two. Following closely is FaceFormer [13], which utilizes long-term audio context through a transformer-based model. Compared to previous methods, the performance of FaceDiffuser [44] is relatively average, only slightly better than VOCA. On the Multiface dataset [19], the results are different. CodeTalker achieved relatively average results. In contrast, FaceDiffuser achieved better results. This may be because the facial regions and dynamics in the MultiFace dataset are more complex. It includes hair regions, more intense movements of the upper face and neck, and actions like blinking, which are not present in the VOCASET. Other methods mainly pay attention to lip-sync and therefore ignore these facial regions. FaceDiffuser uses a diffusion structure to recover

Table 1. Comparison of Lip-Sync Errors (LVE $\times 10^{-4}$), Other Region Errors (OVE $\times 10^{-4}$), and the Upper-FDD ($\times 10^{-5}$)

Dataset	Method (↓)	LVE (↓)	OVE (↓)	Upper-FDD (↓)
VOCASET [7]	VOCA [10]	6.543	4.131	8.190
	FaceFormer [13]	5.280	3.291	4.645
	CodeTalker [56]	4.654	3.211	4.119
	FaceDiffuser [44]	5.591	3.477	4.813
	DiffPoseTalk [46]	4.531	3.029	4.211
	Perceptual 3D Talk [4]	4.472	3.016	4.176
	Ours	**4.423**	**2.941**	**4.107**
Multiface [19]	VOCA [10]	25.041	17.329	14.124
	FaceFormer [13]	6.897	4.438	5.006
	CodeTalker [56]	19.537	13.732	6.674
	FaceDiffuser [44]	6.135	4.071	5.870
	DiffPoseTalk [46]	6.053	3.429	5.002
	Perceptual 3D Talk [4]	5.912	3.376	4.978
	Ours	**5.768**	**3.320**	**4.937**

We compare our method with four methods (VOCA [10], FaceDiffuser [44], FaceFormer [13], CodeTalker [56]) on VOCASET [10] set and Multiface [19] set. Bold text indicates the best performance in that metric.

a face mesh through a denoise procedure, which is more suitable for modeling the facial dynamics. Similarly, on this challenging dataset, DiffPoseTalk and Perceptual 3D Talk demonstrate robust performance, ranking as the closest competitors to our method. This indicates that their underlying architectures—a diffusion-based framework and a perceptually oriented model, respectively—are also effective at handling complex facial dynamics. Our method uses two branches to model the facial region and the mouth region separately, addressing the issue of previous methods that only focused on the mouth region to some extent.

Head Motion Generation. The prediction results for head motion are delineated in Table 2. Our method attains the highest performance. Following closely are two transformer-based approaches: Autoformer [54] and Informer [66], revealing the substantial potential of transformer models in sequence prediction tasks. Methods based on LSTM and GRU models [36] have worse performance. Different from transformers, these methods struggle with single forward step predictions and require iterative generation. During the prediction process, they can only predict the next value one by one in the sequence, so they must predict 50 times to get the final results in our experiments, which introduces a notable amount of errors. Notably, the test results on the RAVDESS dataset consistently have smaller scores compared to those on the VoxCeleb dataset. This disparity can be attributed to the VoxCeleb dataset's increased diversity and complexity in head movements, presenting greater challenges for learning algorithms.

4.3 Qualitative Evaluation

We use the proposed MGM to generate a talking mesh video while keeping head pose still. Then, the proposed HMPM and four baselines are employed to generate a head motion sequence. The generated talking mesh with still heads on VOCASET [7] are illustrated in Figure 4 upper section. As we can see, ours generate a lip shape closest to the Ground Truth. For example, in the first row, the mouth opening in the mesh generated by our method and Ground Truth are larger than

Table 2. The Prediction Accuracy (MSE $\times 10^{-3}$ and MAE $\times 10^{-2}$) of QuaterNet (LSTM) [36], QuaterNet (GRU) [36], Informer [66], Autoformer [54] and Ours in RAVDESS Dataset [31] and VoxCeleb Dataset [7]

Dataset	Method	MSE (↓)	MAE (↓)	DTW Loss (↓)
RAVDESS [31]	QuaterNet (LSTM) [36]	4.751	5.83	11.325
	QuaterNet (GRU) [36]	4.473	5.304	8.572
	Informer [66]	3.256	4.554	3.234
	Autoformer [54]	3.156	4.416	3.022
	Ours	**3.097**	**4.348**	**1.668**
VoxCeleb [7]	QuaterNet (LSTM) [36]	5.012	6.22	12.488
	QuaterNet (GRU) [36]	4.713	5.873	9.858
	Informer [66]	3.653	4.673	4.851
	Autoformer [54]	3.595	4.595	4.535
	Ours	**3.455**	**4.507**	**2.002**

Bold text indicates the best performance in that metric.

meshes generated by other methods. Among the other methods, Perceptual 3D Talk [4] introduces a speech-mesh synchronized representation to meet Temporal Synchronization, Lip Readability, and Expressiveness criteria. However, its fitting capability remains underutilized. In contrast, our approach strategically divides the talking head mesh task into two distinct sub-tasks. Each sub-task is handled by dedicated sub-modules, allowing for the generation of different regions. This division enables our method to achieve superior results compared to Perceptual 3D Talk.

To visualize facial motion dynamics, we refer to CodeTalker [56] to calculate the temporal statistics of adjacent frame facial motions within a sequence. A higher mean value indicates stronger facial movements, while a higher standard deviation suggests a richer variation in facial dynamics. The results shown in the lower section of Figure 4 demonstrate that our method outperforms others in achieving both stronger facial movements and a broader range of dynamics.

On the Multiface [19] (Figure 5), overall, our method is generally comparable to other approaches, particularly FaceDiffuser, with no significant differences. However, our method still slightly outperforms others in terms of lip-sync accuracy. In the first row, the mouth is in the process of closing. Compared to CodeTalker, FaceDiffuser, DiffPoseTalk, and Perceptual 3D Talk, our mouth shape is closer to being closed. FaceFormer and VOCA generate very small mouth movements, almost static, which also results in a closed appearance. In the second and third rows, the mouth is opening, and our results show a wider mouth opening compared to the other methods. As for the facial motion dynamics, ours perform better than other methods.

The generated talking mesh videos with head natural movement are illustrated in Figure 6. We uniformly select 10 frames from the Ground Truth and the generated videos using different methods. To highlight the head motion clearly, we incorporate a heatmap to visualize the differences of both frames, shown in the last row of Figure 6. Natural facial animation videos are expected to showcase prominent and fluid head movements. Notably, in the original video, the head is shaking left and right (indicated by the red box), which is reflected in the boundary area of the face in the heatmap (highlighted by the black box). Our method similarly manifests smooth and natural head movements–gradually looking down (as highlighted by the red box), which is reflected in the nose region in the heatmap (highlighted by the black box).

In contrast, the outputs from Autoformer [54] and Informer [66] depict a face consistently oriented toward the camera, with minimal observable motion. QuarterNet (LSTM) [36] and QuarterNet

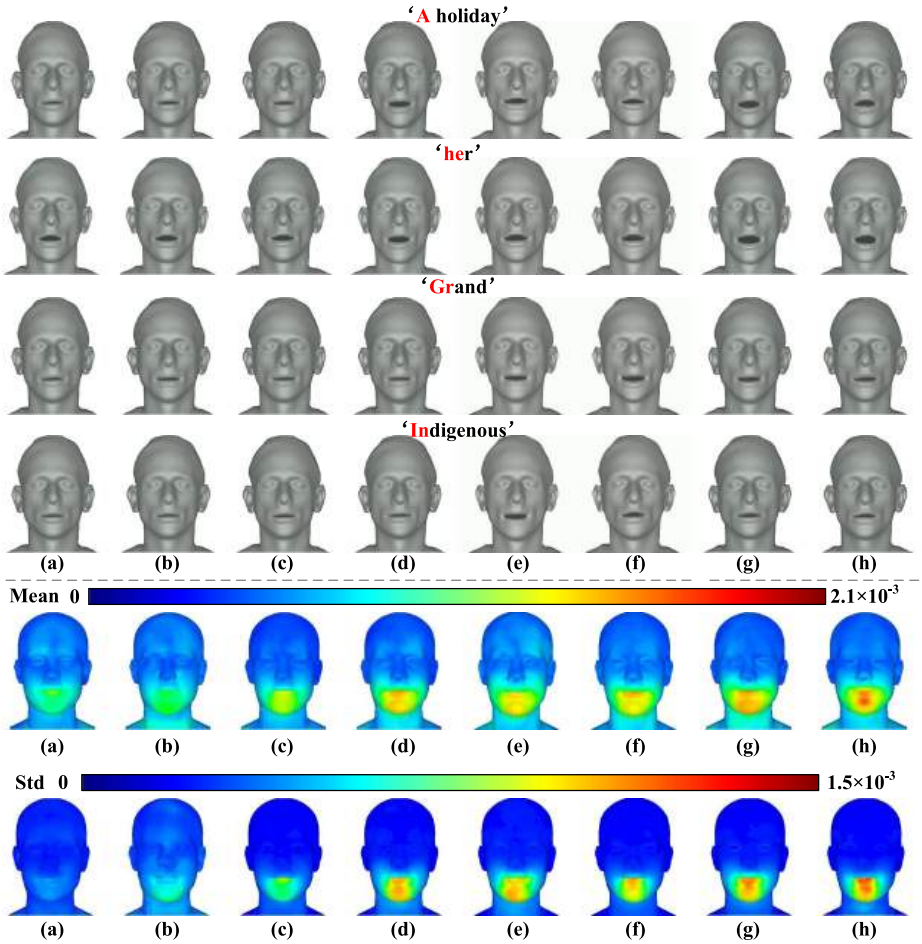

Fig. 4. The visualization of the speech-driven head mesh results on the VOCASET [7], (a) VOCA [10], (b) FaceFormer [13], (c) FaceDiffuser [13], (d) CodeTalker [56], (e) DiffPoseTalk [46], (f) Perceptual 3D Talk [4], (g) Ours, and (h) Ground Truth. The upper section shows the facial animation based on different parts of speech, while the lower section displays the temporal statistics (mean and standard deviation) of motion variations between adjacent frames in a sequence.

(GRU) [36] generate results wherein the head remains static in the same pose, and QuarterNet (LSTM) [36] occasionally exhibits abrupt changes in head position (indicated by the blue box). From the final row of Figure 6, significant deviations in facial contour and nose regions are observed between the heatmaps of the ground truth and our results (notably, darker areas within the black box in the first row and blurred facial areas within the black box in the second row). These discrepancies imply pronounced motion in the heads depicted in these videos. Conversely, alternative methods predominantly manifest heightened values solely within the mouth region, while exhibiting minimal variation elsewhere. This disparity suggests an inadequate representation of natural head movements by these methods.

Subsequently, we proceed to analyze the generated sequence. Specifically, we display a sequence with a length of 250, generated by different methods. To achieve better visual clarity, we added specific offsets to the values of the sequence generated by different methods, staggering them to prevent overlap and ensure clear viewing. The sequences from all methods are depicted in

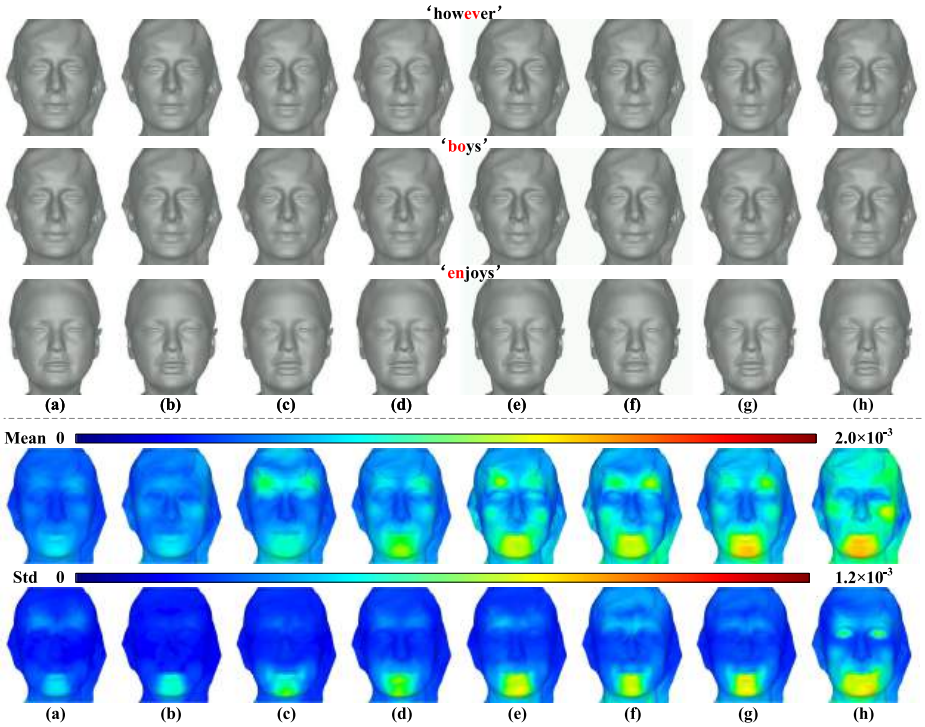

Fig. 5. The visualization of the speech-driven head mesh results on the Multiface [19], (a) VOCA [10] (b) FaceFormer [13], (c) CodeTalker [56], (d) FaceDiffuser [44], (e) DiffPoseTalk [46], (f) Perceptual 3D Talk [4], (g) Ours, and (h) Ground Truth. The upper section shows the facial animation based on different parts of speech, while the lower section displays the temporal statistics (mean and standard deviation) of motion variations between adjacent frames in a sequence.

Figure 7. The outcomes from QuaterNet (LSTM) [36] exhibit considerable unsmooth. Except for the beginning parts, the changes in values are minimal (highlighted by the gray box in the green line). This tendency may cause the generated facial animation video to exhibit some unnatural jitter at the beginning, followed by a tendency to become still. The sequence generated by QuaterNet (GRU) [36] shows very high-frequency fluctuations with large value changes. These drastic value changes (highlighted by the gray bounding box in the blue line) can result in sudden and rough rotations of the head in the generated facial animation video. The Autoformer [54] generates a periodic sequence that appears to repeat the section highlighted by the black box in the red line. This phenomenon is similar to mode collapse in image generation, where each generated sequence follows a fixed pattern. The QuaterNet (LSTM) [36] and Informer [66] generate a relatively stable sequence, with value changes not exceeding 0.02 (highlighted by the gray box in the green and orange lines). This phenomenon results in the head having minimal movement, making it appear almost static in the generated facial animation videos. Besides, we can find that as the length of the generated sequences increases, these methods tend to either produce a sequence that is nearly a straight line (with minimal value changes) or fall into a fixed pattern; however, ours can still generate fluctuating sequences (highlighted by the gray box in the yellow lines). This phenomenon means that our model is *stable* when generating a long sequence.

Additionally, we conducted a visual comparison of existing talking head methods capable of generating head movements, and the results are shown in Figure 8. We selected DiffPoseTalk [46]

Fig. 6. Results of head movement in facial animation video generated by different methods. (a) Original motion sequences, (b) Ours, (c) Autoformer [54], (d) Informer [66], (e) QuaterNet (LSTM) [36], and (f) QuaterNet (GRU) [36]. The last column visualizes the difference between images from the second and fourth columns. We can observe that the heads in the original video and ours have significant movement, while in most of our baselines, the heads remain approximately still. Besides, the result of QuaterNet (LSTM) is unstable. The blue box shows the head suddenly turning to the right and suddenly turning back.

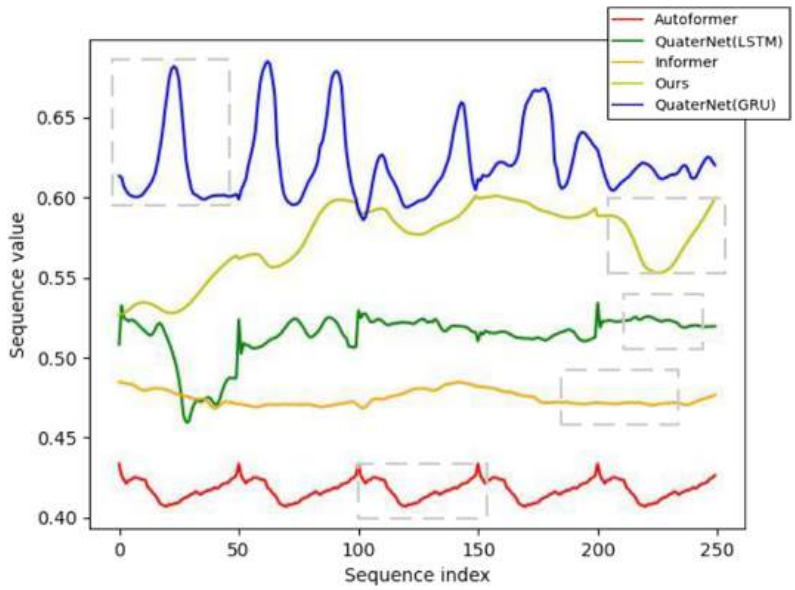

Fig. 7. Visualization of head motion generated by different models.

and SadTalker [64] for comparison, using the same audio clip to generate the results. It can be observed that SadTalker produces a relatively small range of head motion, whereas our method and DiffPoseTalk generate more pronounced and natural head movements.

Fig. 8. Visual results of talking head methods capable of generating head movements, (a) DiffPoseTalk [46] and (b) SadTalker [64].

Table 3. The Results of the Ablation Study about Two Dual Encoders and Corresponding Loss L_M

Dataset	Result	$w/\ L_M$		$w/o\ L_M$	
		LVE (↓)	**OVE (↓)**	**LVE (↓)**	**OVE (↓)**
VOCASET [7]	$w/o\ y_{mouth}$	5.201	3.216	5.141	3.336
	$w/o\ y_{other}$	4.686	3.398	5.136	3.328
	All	**4.423**	**2.941**	**4.527**	**3.056**
Multiface [19]	$w/o\ y_{mouth}$	6.877	3.658	6.767	3.708
	$w/o\ y_{other}$	5.896	3.796	6.763	3.711
	All	**5.768**	**3.320**	**6.247**	**3.431**

Bold text indicates the best performance in that metric.

4.4 Ablation Study

Mesh Generative Model with Dual Encoders. For our mesh generative model, we introduce dual encoders along with corresponding loss functions to specifically target the generation of different mesh regions (y_{mouth} and y_{other}). Here, we conduct ablation studies about the dual encoders and the mouth loss L_M on VOCASET and Multiface. We report LVE and OVE for three variants: adding both y_{mouth} and y_{other}, only adding y_{other} ($w/o\ y_{mouth}$), and only adding y_{mouth} ($w/o\ y_{other}$).

The results are shown in Table 3. When the model is trained with L_M, the mesh $w/o\ y_{mouth}$ has a relatively lower OVE and a relatively higher LVE. When the model is trained $w/o\ L_M$, the two evaluations of two meshes ($w/o\ y_{mouth}$ and $w/o\ y_{other}$) are very close. This demonstrates that L_M makes Encoder_1 focus on generating vertices in the lip region, while Encoder_2 focuses on generating vertices in other areas to complement Encoder_1. Besides, the LVE and OVE of the mesh *all* (adding both y_{mouth} and y_{other}) generated by the model trained using L_M is better than the mesh *all* generated by the model trained without L_M. It means that L_M helps the model to generate more accurate meshes.

For the results on the two datasets, we can see that the values on Multiface are generally higher than those on VOCASET. This is because the meshes in Multiface are far more complex than those in VOCASET, with more intricate facial dynamics, making the model harder to fit. Additionally, on the Multiface dataset, the difference between the model trained without L_M and with L_M is larger than that on VOCASET. This indicates the effectiveness of the strategy of separately generating the mouth region and the face region on complex datasets.

Neural ODEs. We conduct the ablation studies on the Neural ODE of HMPM. The length of the initial head motion state and the Ground Truth sequence is different. The neural ODE not only plays an important role in modeling continuous dynamic sequence but also makes the length of

Table 4. The Results of Ablation Study about Neural ODE on RAVDESS Set and VoxCeleb Set

Dataset	Neural ODE	MSE (↓)	MAE (↓)	DTW Loss (↓)
RAVDESS [31]	w/o neural ODE	3.247	4.571	3.121
	$w/$ **neural ODE**	**3.097**	**4.348**	**1.668**
VoxCeleb [7]	w/o neural ODE	3.617	4.681	4.637
	$w/$ **neural ODE**	**3.455**	**4.507**	**2.002**

Bold text indicates the best performance in that metric.

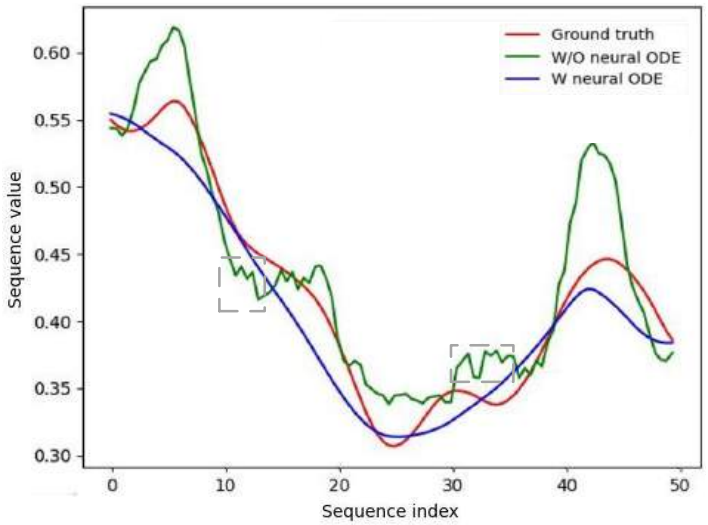

Fig. 9. Ground Truth and the prediction results of the model with neural ODE module ($w/$ neural ODE) and without neural ODE module (w/o neural ODE).

the dynamic sequence equal to the output sequence through using a time sequence t has the same length as the output sequence. If we directly remove the neural ODE, the predicted head motion sequence will not match the Ground Truth. Therefore, we follow Informer [66] and Autoformer [54] to add a zero-valued sequence after the initial head motion state to make the length of the padded sequence equal to the Ground Truth. Table 4 shows the prediction accuracy and change tendency similarity decreased when removing neural ODE, which undoubtedly demonstrates modeling a continuous dynamic sequence can help to improve performance. The prediction results of the HMPM $w/$ neural ODE module and w/o neural ODE module are shown in Figure 9. We observe that the results of the model $w/$ neural ODE are more similar to the Ground Truth. The sequence generated by the model w/o neural ODE has several areas that lack smoothness (as highlighted by the gray boxes), which demonstrates the ability of neural ODE in modeling the continuous dynamic sequence.

4.5 User Study

We conduct user studies on talking mesh generation (keeping head static) and head motion generation. Most of the participants are engaged in the research areas relevant to image generation and talking head. For the comparison of talking mesh generation with head keeping static, we provided

Table 5. The User Study on Talking Mesh Generation (Keeping Head Static)

Dataset	Method	Lip-Sync (↑)	Upper Facial Dynamics (↑)
VOCASET [7]	VOCA [10]	3.35	3.40
	FaceFormer [13]	3.50	4.15
	CodeTalker [56]	4.10	4.35
	FaceDiffuser [44]	3.75	4.20
	Ours	**4.55**	**4.45**
Multiface [19]	VOCA [10]	3.30	3.35
	FaceFormer [13]	3.50	4.05
	CodeTalker [56]	4.05	4.20
	FaceDiffuser [44]	4.25	4.40
	Ours	**4.30**	**4.45**

Bold text indicates the best performance in that metric.

Table 6. The User Study on Head Motion Generation

	Whether Head Movement Is Natural	
	VoxCeleb [7] (↑)	RAVDESS [31] (↑)
QuaterNet (GRU) [36]	1.25	1.20
QuaterNet (LSTM) [36]	1.30	1.20
Informer [66]	1.80	1.95
Autoformer [54]	1.90	2.10
Ours (*w/* noise)	**3.55**	**4.05**
Ours (*w/o* noise)	**3.60**	**4.15**

Bold text indicates the best performance in that metric.

results from our method, four compared methods, and Ground Truth were given to participants. All participants were instructed to rate (0–5) according to two criteria: whether the lip is synchronous with audio (*lip sync*) and whether the upper facial dynamics of the generated meshes is consistent with the Ground Truth (*upper facial dynamics*). The average scores of 20 participants are shown in Table 5. Ours perform better in both *lip sync* and *upper facial dynamics* than other methods. According to participant feedback, our method has a greater advantage over other methods on the VOCASET. Although there is no significant advantage on the Multiface dataset, our method is still slightly better than the others. The Multiface dataset has much richer Ground Truth facial motions compared to the VOCASET. In this context, our method is relatively closer to FaceDiffuser and the Ground Truth. However, all methods struggle to generate blinking actions accurately.

For the user study on head motion generation, we divide them into two groups: one is generated without noise disturbance (*w/o* noise) and the other is generated with noise (*w/* noise). All participants were asked to score according to a standard: whether the head motion is natural (*Natural*). The average scores of 20 participants on two dataset are shown in Table 6. Our score is greatly higher than other methods over all cases. When adding noise disturbance, the score decreases slightly but is still higher than other methods.

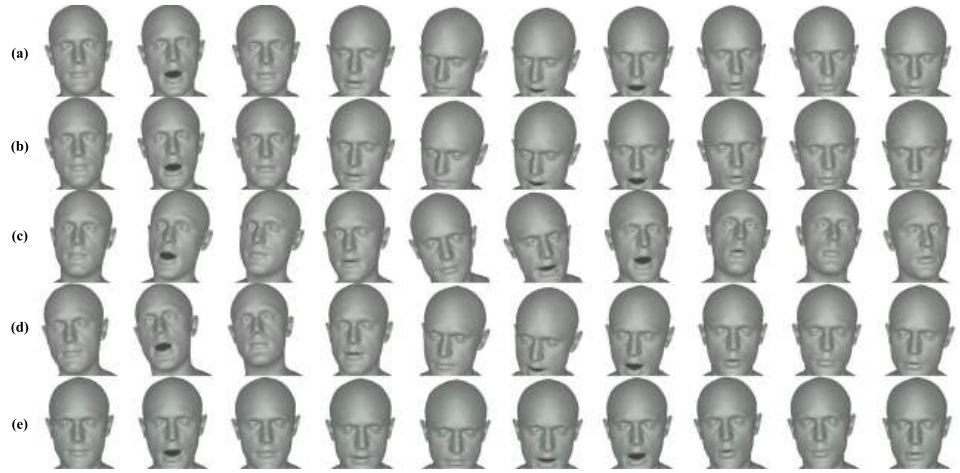

Fig. 10. By adding noise distribution generated by M_N, M_P is capable of generating diverse head motion sequences, allowing for the control of various movements in the head of a facial animation video. Each row corresponds to a different motion sequence.

Table 7. Quantitative Analysis about Diverse Head Motion Generation

Dataset	Noise	DTW Loss (↑)
RAVDESS [31]	w/o noise	1.668
	$w/$noise	9.389
VoxCeleb [7]	w/o noise	2.002
	$w/$ noise	11.2673

4.6 Diverse Head Motion Generation

We also do a diverse head motion generation evaluation. In this experiment, we generate different head motions using the same initial head motion state $S_i \in R^{1\times3}$ and input audio. The visualization results are shown in Figure 10. As we can see, we generate facial animations with different head movements. We also do some quantitative analysis about diverse head motion generation. The DTW loss between the generated head motion sequence is used to evaluate diversity. We generate a head motion sequence without noise (w/o noise) and 10 head motion sequences with noise ($w/$ noise) and then, respectively, calculate DTW losses between them and a head motion sequence extracted from the 2D dataset. All results are shown in Table 7, as we can see, the loss of the sequence without noise is relatively lower which demonstrates that the head motion sequence without noise is similar to the head motion sequence extracted from 2D dataset. When adding noise, the value of DTW loss increases which means that adding noise can generate head motions with different change tendencies. In the user study part (shown in Table 6), we can see that the head motion is still relatively natural when adding noise.

To summarize, our NDM can help generate diverse head motion sequences while keeping head motion natural.

4.7 Limitation and Future Work

While our method achieves accurate lip synchronization, precise vertex positioning, and realistic head motion, generating videos that incorporate emotional variations and subtle facial dynamics (e.g., blinking) remains challenging, limiting overall naturalness. This difficulty stems from the scarcity of emotional talking mesh datasets and the inherently lower expressiveness of 3D meshes compared to 2D images that capture texture-based emotional cues. The absence of fine-grained motions like blinking is partly due to insufficient control mechanisms; some approaches address this via personalized style vectors or one-hot action control in image-based models [57, 64]. Furthermore, our method is primarily designed for natural speech scenarios, and handling exaggerated expressions, rapid head turns, or highly emotional speech remains difficult due to model capacity and training data coverage. The fixed-step Euler discretization in our Neural ODE module, while stable for typical motions, may face numerical challenges with very high–frequency movements.

Additionally, although our model captures rhythmic and statistical patterns of head motion, it does not explicitly model high-level semantic correlations (e.g., nodding for emphasis). Movements are driven mainly by acoustic–prosodic features and personal style, potentially inheriting some implicit semantic associations from the training data but lacking explicit control. This points to an important future direction for achieving more contextually appropriate animations.

In future work, we plan to incorporate richer control information and develop stochastic processes to model subtle facial movements (e.g., blinking). We aim to enhance model capacity for extreme expressions and rapid motions by exploring adaptive numerical solvers and more expressive architectures. Moreover, we intend to integrate multimodal inputs (e.g., text transcripts and semantic labels) to enable explicit semantic control over head and facial movements, thereby improving contextual appropriateness and expressiveness. Extending our framework with larger and more diverse datasets to support a wider range of speech scenarios is another promising direction.

4.8 Conclusion

In this article, we introduce a speech-driven facial animation method that produces natural and consistent head movements while ensuring accurate audio and lip synchronization. Our approach includes an MGM and an HMPM. Compared to earlier speech-driven facial animation methods, our MGM, which incorporates a cascaded mesh generation module, delivers improved results in upper facial dynamics and lip synchronization. Even with datasets that feature complex meshes and intricate facial dynamics, our model performs exceptionally well. Additionally, by using neural ODEs instead of traditional sequence models, our HMPM achieves both accurate predictions and realistic head movements during speech.

References

[1] T. Afouras, J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman. 2018. Deep audio-visual speech recognition. arXiv:1809.02108. Retrieved from https://arxiv.org/abs/1809.02108

[2] Shivangi Aneja, Justus Thies, Angela Dai, and Matthias Nießner. 2024. FaceTalk: Audio-driven motion diffusion for neural parametric head models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 21263–21273.

[3] Joos-Hendrik Böse, Valentin Flunkert, Jan Gasthaus, Tim Januschowski, Dustin Lange, David Salinas, Sebastian Schelter, Matthias Seeger, and Yuyang Wang. 2017. Probabilistic demand forecasting at scale. *Proceedings of the VLDB Endowment* 10, 12 (2017), 1694–1705.

[4] Lee Chae-Yeon, Oh Hyun-Bin, Han EunGi, Kim Sung-Bin, Suekyeong Nam, and Tae-Hyun Oh. 2025. Perceptually accurate 3D talking head generation: New definitions, speech-mesh representation, and evaluation metrics. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 21065–21074.

[5] Ricky T. Q. Chen, Yulia Rubanova, Jesse Bettencourt, and David K. Duvenaud. 2018. Neural ordinary differential equations. In *Advances in Neural Information Processing Systems*, Vol. 31.

[6] Weiqi Chen, Wenwei Wang, Bingqing Peng, Qingsong Wen, Tian Zhou, and Liang Sun. 2022. Learning to rotate: Quaternion transformer for complicated periodical time series forecasting. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '22)*. ACM, New York, NY, 146–156. DOI: https://doi.org/10.1145/3534678.3539234

[7] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. 2018. VoxCeleb2: Deep speaker recognition. arXiv:1806.05622. Retrieved from https://arxiv.org/abs/1806.05622

[8] J. S. Chung and A. Zisserman. 2016. Lip reading in the wild. In *Proceedings of the Asian Conference on Computer Vision*.

[9] Razvan-Gabriel Cirstea, Chenjuan Guo, Bin Yang, Tung Kieu, Xuanyi Dong, and Shirui Pan. 2022. Triformer: Triangular, variable-specific attentions for long sequence multivariate time series forecasting–full version. arXiv:2204.13767. Retrieved from https://arxiv.org/abs/2204.13767

[10] Daniel Cudeiro, Timo Bolkart, Cassidy Laidlaw, Anurag Ranjan, and Michael J. Black. 2019. Capture, learning, and synthesis of 3D speaking styles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10101–10111.

[11] Marco Cuturi and Mathieu Blondel. 2017. Soft-DTW: A differentiable loss function for time-series. In *Proceedings of the International Conference on Machine Learning*. PMLR, 894–903.

[12] Radek Daněček, Michael J. Black, and Timo Bolkart. 2022. EMOCA: Emotion driven monocular face capture and animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20311–20322.

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

[14] Yao Feng, Haiwen Feng, Michael J. Black, and Timo Bolkart. 2021. Learning an animatable detailed 3D face model from in-the-wild images. *ACM Transactions on Graphics* 40, 4 (2021), 1–13.

[15] Simon Giebenhain, Tobias Kirschstein, Markos Georgopoulos, Martin Rünz, Lourdes Agapito, and Matthias Nießner. 2024. MonoNPHM: Dynamic head reconstruction from monocular videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10747–10758.

[16] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial networks. arXiv:1406.2661. Retrieved from https://arxiv.org/abs/1406.2661

[17] Yang Hong, Bo Peng, Haiyao Xiao, Ligang Liu, and Juyong Zhang. 2022. HeadNeRF: A real-time NeRF-based parametric head model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR '22)*.

[18] Xianxu Hou, Xiaokang Zhang, Yudong Li, and Linlin Shen. 2022. TextFace: Text-to-style mapping based face generation and manipulation. *IEEE Transactions on Multimedia* 25 (2022), 3409–3419.

[19] Cheng Hsin Wuu, Ningyuan Zheng, Scott Ardisson, Rohan Bali, Danielle Belko, Eric Brockmeyer, Lucas Evans, Timothy Godisart, Hyowon Ha, Xuhua Huang, et al. 2023. Multiface: A dataset for neural face rendering. arXiv:2207.11243. Retrieved from https://arxiv.org/abs/2207.11243

[20] Lê Minh Ngô, Sezer Karaoğlu and Theo Gevers. 2021. Self-supervised face image manipulation by conditioning GAN on face decomposition. *IEEE Transactions on Multimedia* 24 (2021), 377–385.

[21] Diederik P. Kingma and Max Welling. 2022. Auto-Encoding Variational Bayes. arXiv:1312.6114. Retrieved from https://arxiv.org/abs/1312.6114

[22] Tobias Kirschstein, Shenhan Qian, Simon Giebenhain, Tim Walter, and Matthias Nießner. 2023. NeRSemble: Multi-view radiance field reconstruction of human heads. *ACM Transactions on Graphics* 42, 4 (2023), 1–14.

[23] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. arXiv:1910.13461. Retrieved from https://arxiv.org/abs/1910.13461

[24] Jiaman Li, Karen Liu, and Jiajun Wu. 2023. Ego-body pose estimation via ego-head pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17142–17151.

[25] Jiahe Li, Jiawei Zhang, Xiao Bai, Jin Zheng, Xin Ning, Jun Zhou, and Lin Gu. 2024. TalkingGaussian: Structure-persistent 3D talking head synthesis via Gaussian splatting. arXiv:2404.15264. Retrieved from https://arxiv.org/abs/2404.15264

[26] Tianye Li, Timo Bolkart, Michael J. Black, Hao Li, and Javier Romero. 2017. Learning a model of facial shape and expression from 4D scans. *ACM Transactions on Graphics* 36, 6 (2017), 194.

[27] Yingjian Li, Zheng Zhang, Bingzhi Chen, Guangming Lu, and David Zhang. 2022. Deep margin-sensitive representation learning for cross-domain facial expression recognition. *IEEE Transactions on Multimedia* 25 (2022), 1359–1373.

[28] Bryan Lim, Sercan Ö. Arık, Nicolas Loeff, and Tomas Pfister. 2021. Temporal fusion transformers for interpretable multi-horizon time series forecasting. *International Journal of Forecasting* 37, 4 (2021), 1748–1764.

[29] Renshuai Liu, Yao Cheng, Sifei Huang, Chengyang Li, and Xuan Cheng. 2023. Transformer-based high-fidelity facial displacement completion for detailed 3D face reconstruction. *IEEE Transactions on Multimedia* 26 (2023), 799–810.

[30] Si Liu, Renda Bao, Defa Zhu, Shaofei Huang, Qiong Yan, Liang Lin, and Chao Dong. 2022. Fine-grained face editing via personalized spatial-aware affine modulation. *IEEE Transactions on Multimedia* 25 (2022), 4213–4224.

[31] Steven R. Livingstone and Frank A. Russo. 2018. 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 (2018), e0196391.

[32] Zhiyuan Ma, Xiangyu Zhu, Guo-Jun Qi, Zhen Lei, and Lei Zhang. 2023. OTAvatar: One-shot talking face avatar with controllable tri-plane rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16901–16910.

[33] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. 2020. NeRF: Representing scenes as neural radiance fields for view synthesis. arXiv:2003.08934. Retrieved from https://arxiv.org/abs/2003.08934

[34] Manfred Mudelsee. 2019. Trend analysis of climate time series: A review of methods. *Earth-Science Reviews* 190 (2019), 310–322.

[35] Arsha Nagrani, Joon Son Chung, Weidi Xie, and Andrew Zisserman. 2020. VoxCeleb: Large-scale speaker verification in the wild. *Computer Speech & Language* 60 (2020), 101027.

[36] Dario Pavllo, David Grangier, and Michael Auli. 2018. QuaterNet: A quaternion-based recurrent model for human motion. arXiv:1805.06485. Retrieved from https://arxiv.org/abs/1805.06485

[37] Xiaogang Peng, Siyuan Mao, and Zizhao Wu. 2023. Trajectory-aware body interaction transformer for multi-person pose forecasting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17121–17130.

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

[39] Shenhan Qian, Tobias Kirschstein, Liam Schoneveld, Davide Davoli, Simon Giebenhain, and Matthias Nießner. 2024. GaussianAvatars: Photorealistic head avatars with rigged 3D Gaussians. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20299–20309.

[40] Alexander Richard, Michael Zollhöfer, Yandong Wen, Fernando De la Torre, and Yaser Sheikh. 2021. MeshTalk: 3D face animation from speech using cross-modality disentanglement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1173–1182.

[41] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10684–10695.

[42] Amin Shabani, Amir Abdi, Lili Meng, and Tristan Sylvain. 2023. Scaleformer: Iterative multi-scale refining transformers for time series forecasting. arXiv:2206.04038. Retrieved from https://arxiv.org/abs/2206.04038

[43] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. 2019. First order motion model for image animation. In *Advances in Neural Information Processing Systems*, Vol. 32.

[44] Stefan Stan, Kazi Injamamul Haque, and Zerrin Yumak. 2023. FaceDiffuser: Speech-driven 3D facial animation synthesis using diffusion. arXiv:2309.11306. Retrieved from https://arxiv.org/abs/2309.11306

[45] Fan-Keng Sun and Duane S. Boning. 2022. FreDo: Frequency domain-based long-term time series forecasting. arXiv:220512301. Retrieved from https://arxiv.org/abs/2205.12301

[46] Zhiyao Sun, Tian Lv, Sheng Ye, Matthieu Gaetan Lin, Jenny Sheng, Yu-Hui Wen, Minjing Yu, and Yong-Jin Liu. 2023. DiffPoseTalk: Speech-driven stylistic 3D facial animation and head pose generation via diffusion models. arXiv:2310.00434. Retrieved from https://arxiv.org/abs/2310.00434

[47] Balamurugan Thambiraja, Ikhsanul Habibie, Sadegh Aliakbarian, Darren Cosker, Christian Theobalt, and Justus Thies. 2023. Imitator: Personalized speech-driven 3D facial animation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 20621–20631.

[48] Eric J. Topol. 2019. High-performance medicine: The convergence of human and artificial intelligence. *Nature Medicine* 25, 1 (2019), 44–56.

[49] Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. 2018. Neural discrete representation learning. arXiv:171100937. Retrieved from https://arxiv.org/abs/1711.00937

[50] Jiacheng Wang, Ping Liu, Jingen Liu, and Wei Xu. 2023. Text-guided eyeglasses manipulation with spatial constraints. *IEEE Transactions on Multimedia* 26 (2023), 4375–4388.

[51] Qianyun Wang, Zhenfeng Fan, and Shihong Xia. 2021. 3D-TalkEmo: Learning to synthesize 3D emotional talking head. arXiv:2104.12051. Retrieved from https://arxiv.org/abs/2104.12051

[52] Yuyang Wang, Alex Smola, Danielle Maddix, Jan Gasthaus, Dean Foster, and Tim Januschowski. 2019. Deep factors for forecasting. In *Proceedings of the International Conference on Machine Learning*. PMLR, 6607–6617.

[53] Dong Wei, Huaijiang Sun, Bin Li, Jianfeng Lu, Weiqing Li, Xiaoning Sun, and Shengxiang Hu. 2023. Human joint kinematics diffusion-refinement for stochastic motion prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37, 6110–6118.

[54] Haixu Wu, Jiehui Xu, Jianmin Wang, and Mingsheng Long. 2021. Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. In *Advances in Neural Information Processing Systems*, Vol. 34, 22419–22430.

[55] Weicheng Xie, Wenya Lu, Zhibin Peng, and Linlin Shen. 2023. Consistency preservation and feature entropy regularization for GAN based face editing. *IEEE Transactions on Multimedia* 25 (2023), 8892–8905.

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

[57] Mingwang Xu, Hui Li, Qingkun Su, Hanlin Shang, Liwei Zhang, Ce Liu, Jingdong Wang, Luc Van Gool, Yao Yao, and Siyu Zhu. 2024. Hallo: Hierarchical audio-driven visual synthesis for portrait image animation. arXiv:2406.08801. Retrieved from https://arxiv.org/abs/2406.08801

[58] Yuelang Xu, Benwang Chen, Zhe Li, Hongwen Zhang, Lizhen Wang, Zerong Zheng, and Yebin Liu. 2024. Gaussian head avatar: Ultra high-fidelity head avatar via dynamic Gaussians. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1931–1941.

[59] Wenwu Yang, Yeqing Zhao, Bailin Yang, and Jianbing Shen. 2023. Learning 3D face reconstruction from the cycle-consistency of dynamic faces. *IEEE Transactions on Multimedia* 26 (2023), 3663–3675.

[60] Rajeev Yasarla, Federico Perazzi, and Vishal M. Patel. 2020. Deblurring face images using uncertainty guided multi-stream semantic networks. *IEEE Transactions on Image Processing* 29 (2020), 6251–6263.

[61] Zhenhui Ye, Ziyue Jiang, Yi Ren, Jinglin Liu, JinZheng He, and Zhou Zhao. 2023. GeneFace: Generalized and high-fidelity audio-driven 3D talking face synthesis. arXiv:2301.13430. Retrieved from https://arxiv.org/abs/2301.13430

[62] Tom Young, Devamanyu Hazarika, Soujanya Poria, and Erik Cambria. 2018. Recent trends in deep learning based natural language processing. *IEEE Computational Intelligence Magazine* 13, 3 (2018), 55–75.

[63] Lingyun Yu, Hongtao Xie, and Yongdong Zhang. 2021. Multimodal learning for temporally coherent talking face generation with articulator synergy. *IEEE Transactions on Multimedia* 24 (2021), 2950–2962.

[64] Wenxuan Zhang, Xiaodong Cun, Xuan Wang, Yong Zhang, Xi Shen, Yu Guo, Ying Shan, and Fei Wang. 2023. 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*, 8652–8661.

[65] Qingkai Zhen, Di Huang, Yunhong Wang, and Liming Chen. 2016. Muscular movement model-based automatic 3D/4D facial expression recognition. *IEEE Transactions on Multimedia* 18, 7 (2016), 1438–1450.

[66] Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang. 2021. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35, 11106–11115.

[67] Tian Zhou, Ziqing Ma, Qingsong Wen, Xue Wang, Liang Sun, and Rong Jin. 2022. FEDformer: Frequency enhanced decomposed transformer for long-term series forecasting. In *Proceedings of the 39th International Conference on Machine Learning*, Vol. 162. Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (Eds.), PMLR, 27268–27286. Retrieved from https://proceedings.mlr.press/v162/zhou22g.html

Received 8 May 2025; revised 3 February 2026; accepted 14 February 2026