KMTalk: Speech-Driven 3D Facial Animation with Key Motion Embedding

Zhihao Xu[1*], Shengjie Gong[1*], Jiapeng Tang[2], Lingyu Liang[1], Yining Huang[1], Haojie Li[1], and Shuangping Huang[1,3†]

[1] South China University of Technology
[2] Technical University of Munich
[3] Pazhou Laboratory
{eezhihaoxu,eeshengjiegong}@mail.scut.edu.cn, eehsp@scut.edu.cn

Abstract. We present a novel approach for synthesizing 3D facial motions from audio sequences using key motion embeddings. Despite recent advancements in data-driven techniques, accurately mapping between audio signals and 3D facial meshes remains challenging. Direct regression of the entire sequence often leads to over-smoothed results due to the ill-posed nature of the problem. To this end, we propose a progressive learning mechanism that generates 3D facial animations by introducing key motion capture to decrease cross-modal mapping uncertainty and learning complexity. Concretely, our method integrates linguistic and data-driven priors through two modules: the linguistic-based key motion acquisition and the cross-modal motion completion. The former identifies key motions and learns the associated 3D facial expressions, ensuring accurate lip-speech synchronization. The latter extends key motions into a full sequence of 3D talking faces guided by audio features, improving temporal coherence and audio-visual consistency. Extensive experimental comparisons against existing state-of-the-art methods demonstrate the superiority of our approach in generating more vivid and consistent talking face animations. Consistent enhancements in results through the integration of our proposed learning scheme with existing methods underscore the efficacy of our approach.

Keywords: Speech-driven · 3D Facial Animation · Key Motion

1 Introduction

Speech-driven 3D facial animation aims to create realistic talking heads that synchronize with input speech. It plays a significant role in many applications of virtual reality, like film production, computer gaming, and education [24,43].

The main challenge of speech-driven 3D talking faces lies in the ill-posed problem caused by the cross-modal mapping uncertainty from the speech domain to

*Authors contributed equally.
†Corresponding author.

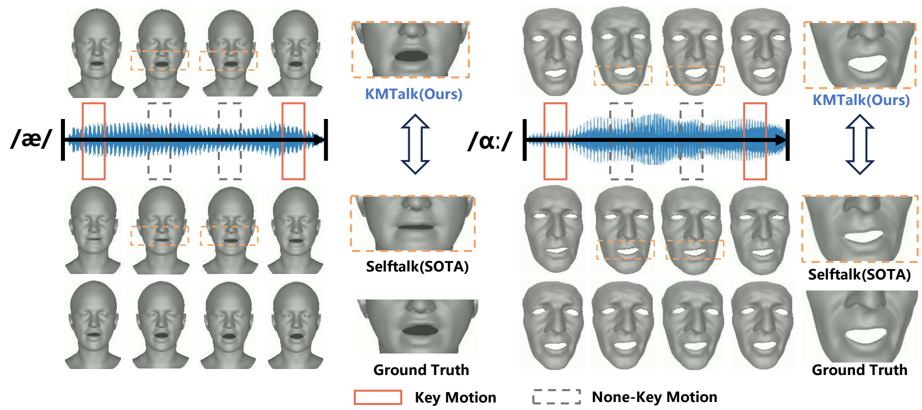


Fig. 1: Compared to the state-of-the-art method Selftalk, our approach can produce more vivid lip motions from speeches, since we introduce linguistic priors to characterize key motions and utilize data-driven priors to interpolate non-key motions.

the 3D motion domain. Since there may be multiple plausible outputs for input audio, effective regularizations and constraints should be integrated into the system, to generate vivid facial motions. The related methods can be roughly divided into linguistic-based methods [3, 11, 29, 46, 54] and data-driven methods [8,12,16,26,34,36,41,49,52,53]. For linguistic-based methods [3,11,29,46,54], a set of intricate phoneme-to-viseme mapping rules is manually designed to generate the talking mouth based on priors from visemes or linguistic knowledge. While these methods explicitly control the animation of articulation processes, such as procedural lip sync with animation curves, their focus is mainly on localized facial movements, like those of the mouth area, lacking a systematic approach for modeling comprehensive facial motion. Thanks to the established audio-to-face datasets, learning-based methods [12,16,34,36,41,49,52,53] choose to map audio signals into 3D facial meshes in a data-driven manner. Most of these works [8, 12, 16, 26, 34, 36, 41, 48, 49, 52, 53] typically formulated the cross-modal mapping of 3D talking face generation as a regression task, such as MeshTalk [41], FaceFormer [12], and SelfTalk [36]. While achieving impressive performance, they exhibit common limitations in their learning schemes. Firstly, they directly learn the ambiguous cross-modal mapping between audio and facial expression sequences, always leading to sub-optimal results in terms of temporal coherence and audio-visual consistency. Secondly, these methods typically regress the entire sequence without considering key motion cues, hindering the capture of detailed facial dynamics and accurate lip movements, particularly in complex facial expressions such as puckering or opening the mouth (as depicted in Fig. 1). Lastly, they overlook linguistic priors essential for simulating the articulation process, thereby limiting their ability to achieve precise lip-speech synchronization.

To this end, inspired by the keyframe-based video generation techniques observed in recent studies [27, 33, 51, 56], which prioritize the generation of keyframes before adding detailed elements, we introduce a progressive learning mechanism that generates realistic 3D facial animations from audio inputs by incorporating key motion embeddings. The key idea is to initially generate key facial expressions, and then interpolate the intermediate motions to obtain the entire motion sequence, which significantly reduces the uncertainty of cross-modal mapping and eases the learning difficulty. Concretely, our method integrates linguistic and data-driven priors through two modules: the linguistic-based key motion acquisition and the cross-modal motion completion. The linguistic-based key motion acquisition module utilizes phoneme-based localization methods to identify temporal indices of key motion, which correspond to significant motion snapshots aligned with phoneme changes in the audio. Once the key motion indexes are determined, a key motion decoder interprets associated 3D facial meshes from corresponding audio features. This highlights those distinct facial expressions and facilitates lip-speech synchronization. The cross-modal motion completion module expands non-continuous key motions into a full sequence of continuous face motions using audio features as guidance. This process enhances audio-mesh alignment and improves the temporal smoothness of output facial mesh sequences. The contributions of our work are summarized as follows:

– We propose a progressive learning mechanism to generate speech-driven 3D talking faces. It uses linguistic priors to initially generate key motions, and then interpolate key motions into complete motions via data-driven priors.
– We propose the use of phoneme-based localization methods to capture key facial motions. It effectively captures significant expression transitions aligned with phoneme changes in audio, improving lip-speech synchronization.
– We design a cross-modal facial motion completion module to produce full sequences of 3D talking faces using synthesized key motions and audio features. It further enhances lip-speech synchronization accuracy while facilitating temporal coherence in facial motions.

Extensive experimental comparisons on the BIWI [13] and VOCASET [8] datasets demonstrate that our method outperforms existing state-of-the-art approaches in more accurate and realistic talking face generation. Detailed ablation studies confirm the effectiveness of our proposed key motion capture technique. Additionally, consistent improvements in results by combining our proposed scheme with existing methods validate the efficacy of our design. Our code and weights will be at the project website: https://github.com/ffxzh/KMTalk.

2 Related Work

While existing research [1,5,6,10,15,18,19,23,25,35,38,42,47,50,55,59,61] focuses on 2D talking heads, we focus on audio-driven 3D facial animations in this work, which can be roughly categorized into linguistics-based and data-driven methods.

2.1 Linguistic-based Methods

Linguistic-based methods [3, 7, 11, 14, 22, 29, 30, 46, 54] establish a set of intricate phoneme-to-viseme mapping rules for animating the mouth. For example, the dynamic viseme model proposed by Taylor *et al*. [46] exploits the one-to-many mapping of phonemes to lip motions. JALI [11] considers the many-to-many mapping between phonemes and visemes. More recently, Bao *et al*. [3] introduced a novel parameterized viseme fitting algorithm that extracts viseme parameters from speech videos using phonemic priors. Leveraging linguistic priors, these methods indicate the articulation process by providing animators with explicit control over animation, thus boosting their performance in lip-speech synchronization. However, these approaches primarily focus on animating the lip region, lacking a comprehensive strategy for animating the entire face. In our work, we leverage linguistic priors to detect key frames with significant expression changes from audio in an analytical manner, without the need for supervised training.

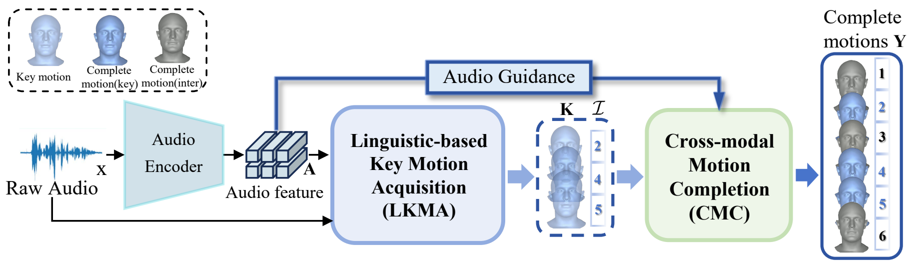
2.2 Data-driven Methods

With the development of deep learning technology and the availability of high-quality datasets, data-driven methods [4, 8, 12, 16, 20, 26, 34, 36, 37, 41, 44, 45, 49, 52, 53, 60] were proposed to synthesize entire 3D facial animation. Some methods [8,12,20,40,52] attempt to establish a direct audio-to-visual mapping through regression. Person-specific approaches [20, 40] can usually obtain plausible facial motions because of the relatively consistent talking style. VOCA [8] incorporates a robust audio feature extraction model capable of capturing various speaking styles, which can generate realistic speaker-independent animation and shows its wide applicability. MeshTalk [41] constructed a categorical latent space to adaptively generate motions based on the separated audio-correlated and audio-uncorrelated facial information. FaceFormer [12] introduced two biased attention mechanisms and integrated the self-supervised pre-trained speech representations for the ill-posed and data scarcity issues. CodeTalker [53] proposed the discrete motion prior which regards the cross-modal mapping as a code query task in a finite proxy space of the learned codebook. SelfTalk [36] proposed a self-supervised approach to construct a lip-reading interpreter and speech recognizer to enhance the comprehensibility of generated lip movements. While data-driven approaches have shown impressive performance, accurately learning cross-modal audio-visual mappings remains challenging due to inherent uncertainties. These methods often regress the entire sequence, leading to over-smoothing and a lack of detailed facial dynamics. In contrast, our approach employs a coarse-to-fine learning mechanism that separates the problem into key motion capture and motion completion stages. This approach effectively mitigates cross-modal mapping uncertainties and reduces learning complexity, resulting in more precise and dynamic facial animation synthesis.

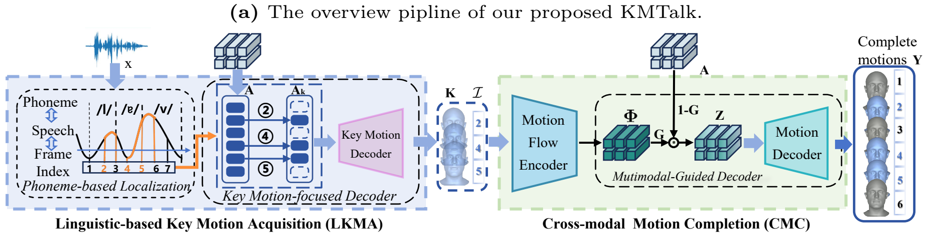
3 Method

3.1 Overview

Let $\mathbf{x}$ represents the raw audio input and $\hat{\mathbf{Y}} = (\hat{\mathbf{y}}_1, ..., \hat{\mathbf{y}}_N) \in \mathbb{R}^{N \times V \times 3}$ denotes the corresponding ground-truth sequence of facial movement over a neutral template, where N indicates the number of visual frames and V denotes the number



(a) The overview pipline of our proposed KMTalk.



(b) The details of two key modules in KMTalk.

Fig. 2: Fig. 2a illustrates the overview pipeline of our proposed KMTalk. Initially, the Audio Encoder takes the input raw audio $\mathbf{x}$ and encodes it into audio features $\mathbf{A}$. Subsequently, in the LKMA module, key motions $\mathbf{K}$ are generated from the audio $\mathbf{x}$ and $\mathbf{A}$. Finally, the CMC module reintroduces audio features $\mathbf{A}$ to extend these key motions $\mathbf{K}$ into a full sequence $\mathbf{Y}$. Fig. 2b presents the details of two key modules in KMTalk. In the Linguistic-based Key Motion Acquisition, a Phoneme-based Localization Method is used to identify key motion indices $\mathcal{I}$ from raw audio $\mathbf{x}$. Based on audio features $\mathbf{A}$ and $\mathcal{I}$, the Key Motion-focused Decoder generates key motions $\mathbf{K}$. In the Cross-modal Motion Completion, the Motion Flow Encoder processes $\mathbf{K}$ and $\mathcal{I}$, producing motion flow features $\mathbf{\Phi}$. Then, with the dynamic fusion weight $\mathbf{G}$, the Multimodal-Guided Decoder combines $\mathbf{\Phi}$ and $\mathbf{A}$ to decode the final motion sequence $\mathbf{Y}$.

of vertices in the facial mesh. The objective is to synthesize $\mathbf{Y} = (\mathbf{y}_1, ..., \mathbf{y}_N)$ that is similar to $\hat{\mathbf{Y}}$, driven by the raw audio $\mathbf{x}$. The generated sequence should ensure lip synchronization with the audio while exhibiting natural facial movements.

Due to the domain gap between modalities and the ill-posed nature of directly translating audio to facial movement sequences, it is a challenging task that often results in over-smooth or poorly synchronized lip movements. To address these issues, this paper introduces a coarse-to-fine approach with key motion embedding, integrating both linguistic and data-driven priors. The overview pipeline is presented in Fig. 2a. In the LKMA module, linguistic priors are introduced to locate and generate higher-quality key motions (see Sec. 3.2), followed by the CMC module, where these key motions are fleshed out into a complete sequence of facial motions (see Sec. 3.3).

3.2 Linguistic-based Key Motion Acquisition

In the realm of audio-driven 3D facial animation, it presents significant challenges to precisely define which frames constitute key motions. An alternative and simple solution is to use uniform or random sampling to determine the positions of these key motions. Although these approaches can boost performance to a certain degree due to the reduced learning complexity (refer to Sec. 4.2 for ablation studies), they fail to utilize the correlation between audio content and facial movements. However, we can leverage linguistic priors to capture pronounced articulatory actions, which are identifiable at phoneme boundaries. This approach circumvents the issue of over-smooth in the output sequence, thereby enhancing the overall quality of the results.

As shown in the left of Fig. 2b, the Linguistic-based Key Motion Acquisition (LKMA) module receives as inputs the raw audio $\mathbf{x}$ and the audio features $\mathbf{A} = (\mathbf{a}_1, ..., \mathbf{a}_N) \in \mathbb{R}^{N\times d}$, where the audio features are derived from the Audio Encoder that utilizes the wav2vec 2.0 pre-trained model [2], processing the raw audio $\mathbf{x}$ as its input. Then it takes raw audio $\mathbf{x}$ as input to produce the key motion indices $\mathcal{I} = \{i_1, ..., i_m\}$, where $i_j \in \{1, ..., N\}, \forall i_j \in \mathcal{I}$, through the proposed Phoneme-based Localization method. Subsequently, the Key Motion-focused Decoder utilizes the audio features $\mathbf{A}$ and the key motion indices $\mathcal{I}$ to generate key motions $\mathbf{K} = (\mathbf{k}_{i_1}, ..., \mathbf{k}_{i_m}) \in \mathbb{R}^{m\times V\times 3}$ consisting of m frames of facial movement which are located on the key motion positions, where $\mathbf{k}_{i_j} \simeq \hat{\mathbf{y}}_{i_j}, \forall i_j \in \mathcal{I}$. The process of the LKMA module is expressed as:

$$\mathcal{I}, \mathbf{K} = \mathrm{LKMA}(\mathbf{x}, \mathbf{A}). \tag{1}$$

Phoneme-based Localization. At phoneme boundaries, a noticeable offset is observed in the articulator movement, with visualization results available in supplementary materials. Furthermore, the phoneme boundary effects underscore the ease with which the boundaries of phonemes can be perceived [17]. Both experimental and theoretical analyses have demonstrated a distinct position-mapping relationship between the phoneme boundaries in the audio and the significant elements in the motion sequence, specifically the key motions. The mapping relationship can be harnessed to facilitate the initial alignment between the audio and visual modalities.

Specifically, an Automatic Speech Recognition model [28] is first utilized to obtain the text content from the raw audio $\mathbf{x}$. Then, a Montreal Forced Aligner [31] is adopted to align the audio and the text, producing the start and the end timestamps for each phoneme. Finally, the indices of these motion frames corresponding to the timestamps are regarded as key motion indices $\mathcal{I}$, and the corresponding motions compose the key motions $\mathbf{K}$.

Key Motion-focused Decoder. It is utilized to synthesize key motions $\mathbf{K}$ of superior quality. Initially, employing $\mathcal{I}$ as indices, we extract the corresponding aligned audio features $\mathbf{A}_\mathbf{k} = (\mathbf{a}_{i_1}, ..., \mathbf{a}_{i_m}) \in \mathbb{R}^{m\times d}$ from the comprehensive audio features $\mathbf{A}$. Subsequently, it adopts a modified transformer-based architecture, processes $\mathbf{A}_\mathbf{k}$ to generate the key motions $\mathbf{K}$.

Loss Function. Intuitively, a straightforward approach to optimize the key motions $\mathbf{K}$ involves utilizing $\mathcal{I}$ to index $\hat{\mathbf{Y}}$, resulting in $\hat{\mathbf{Y}}_\mathbf{k} = (\hat{\mathbf{y}}_{i_1}, \ldots, \hat{\mathbf{y}}_{i_m})$, which serves as the supervision for the training process. However, key motions typically occupy non-adjacent positions within the entire sequence. Hence, given the lack of inter-frame contextual information, attempting direct frame-by-frame regression of $\mathbf{K}$ towards $\hat{\mathbf{Y}}_\mathbf{k}$ may fall short in achieving accurate facial expressions, as well as producing smooth and realistic animation.

To address this limitation, we adopt a pseudo-complete sequence training method that utilizes the ground-truth frame labels $\hat{\mathbf{Y}}$ at the non-key indices $\mathcal{I}' = \{1, ..., N\} \setminus \mathcal{I}$ and the generated key motions $\mathbf{K}$ at the key indices $\mathcal{I}$ to form a predicted **p**seudo-complete sequence $\mathbf{Y_p} \in \mathbb{R}^{N \times V \times 3}$. Then, the model is trained by minimizing the loss between the pseudo-complete sequence $\mathbf{Y_p}$ and the ground-truth sequence $\hat{\mathbf{Y}}$. This enables the model to capture subtle changes between key motions and adjacent ground-truth frames, thereby mitigating inter-frame jitter and achieving more accurate regression of facial expressions. Following the SelfTalk [36], the loss function is formulated as:

$$\mathcal{L}_{LKMA} = \lambda_1 \mathcal{L}_{rec} + \lambda_2 \mathcal{L}_{vel} + \lambda_3 \mathcal{L}_{lat} + \lambda_4 \mathcal{L}_{ctc}, \tag{2}$$

where $\lambda_1 = 1000.0, \lambda_2 = 1000.0, \lambda_3 = 0.001$, and $\lambda_4 = 0.0001$ in all of our experiments. $\mathcal{L}_{rec}$, $\mathcal{L}_{vel}$, and $\mathcal{L}_{lat}$ are measured by mean square error, while $\mathcal{L}_{ctc}$ is quantified by CTC Loss. The reconstruction loss $\mathcal{L}_{rec}$ quantifies the discrepancy between the predicted and the ground-truth facial movements. The velocity loss $\mathcal{L}_{vel}$ reduces frame jitter, ensuring smooth and natural lip movements. The latent consistency loss $\mathcal{L}_{lat}$ assesses the variance between latent features extracted from both the audio and lip shape encoders, aiming to align the learned audio and lip features. Lastly, the text consistency loss $\mathcal{L}_{ctc}$ evaluates the difference between the lip-reading decoder's output and the original text, ensuring the intelligibility of the lip-reading results.

3.3 Cross-modal Motion Completion

A straightforward method to obtain a complete talking face sequence is to directly integrate key motions into the entire face movements. However, it is important to note that while key motions capture essential facial dynamics, they may not encompass all details of non-key motions. This direct integration could result in mismatches between augmented non-key motions and their corresponding audio segments (see Sec. 4.2 for ablation studies). To mitigate this issue, we introduce a Cross-modal Motion Completion (CMC) module that jointly combines the audio features $\mathbf{A}$, key motions $\mathbf{K}$, and key motion indices $\mathcal{I}$ to generate a complete sequence of 3D facial meshes $\mathbf{Y}$. The process can be formulated as:

$$\mathbf{Y} = \mathrm{CMC}(\mathbf{A}, \mathbf{K}, \mathcal{I}). \tag{3}$$

The details of the CMC module are illustrated on the right of Fig. 2b.

Motion Flow Encoder. Key motions serve as a kinematic prior for the remaining frames, offering valuable insights into facial dynamics. To effectively capture the motion flow information provided by key motions, we draw inspiration from some manifold methods [32] to acquire motion flow features $\mathbf{\Phi} \in \mathbb{R}^{N \times d}$ from the key motions $\mathbf{K}$. Specifically, we first encode the key motions $\mathbf{K}$ into key motion context tokens $\mathbf{\Phi_k} \in \mathbb{R}^{m \times d}$ by multiple transformer-encoder layers. At the same time, the indices of non-key motions $\mathcal{I}'$ are encoded into positional encodings by a sinusoidal position embedding layer, representing the non-key frame positions. Then, we adopt the cross-attention layers to extract the intermediate tokens $\mathbf{\Phi_{non\text{-}key}} \in \mathbb{R}^{(N-m) \times d}$, with key and value from linear transformations of $\mathbf{\Phi_k}$ and the query is the positional encodings of non-key frames indices. Above all, the implicit motion manifold proposed in CITL [32] is utilized to arrange $\mathbf{\Phi_k}$ and $\mathbf{\Phi_{non\text{-}key}}$ based on their indices, followed by a 1D convolution for fusion, ultimately obtaining the motion flow features $\mathbf{\Phi}$.

Multimodal-Guided Decoder. The non-key motions' features in the complete motion sequence feature estimation are derived from the global context interpolation of key motions, which has a certain degree of information loss due to the audio feature selection process in the Key Motion-focused Decoder (Sec. 3.2). Hence, we have devised a multimodal-guided decoding approach that incorporates audio modalities to furnish comprehensive information across the entire temporal scale, alongside motion flow to offer facial motion priors. These elements serve to guide and constrain the decoding process, thereby facilitating the precise generation of the motion sequence. Technically, we simply employ the gated mechanism [9, 57, 58] for the modality fusion, which can be formulated as:

$$\mathbf{G} = \sigma([\mathbf{A}, \mathbf{\Phi}]\mathbf{W}), \tag{4}$$

$$\mathbf{Z} = \mathbf{G} \odot \mathbf{A} + (1 - \mathbf{G}) \odot \mathbf{\Phi}, \tag{5}$$

where $\odot$ represents the element-wise multiplication operation, $[\cdot, \cdot]$ denotes the concatenation operation, $\mathbf{W} \in \mathbb{R}^{2d \times d}$ is a parameter and σ is a sigmoid function, while $\mathbf{G} \in \mathbb{R}^{N \times d}$ dynamically selects features from the audio features $\mathbf{A}$ and the motion flow features $\mathbf{\Phi}$. Then, the Motion Decoder, a transformer-based structural model, is employed to transform the fusion features $\mathbf{Z}$ into the complete 3D facial movement sequence $\hat{\mathbf{Y}}$.

Loss Function. We train the CMC module utilizing the reconstruction loss and velocity loss. The total loss function is defined as:

$$\mathcal{L}_{CMC} = \mathcal{L}_{rec} + \mathcal{L}_{vel}. \tag{6}$$

4 Experiment

Dataset. *BIWI* [13] consists of 40 paired audio-visual sentences from 14 subjects. The 3D facial geometries, consisting of 23370 vertices, were captured at a frame rate of 25fps, and the average duration of each sequence was 4.67 seconds. We adopt the same evaluation protocol as FaceFormer [12] on the BIWI dataset. Specifically, the training set (BIWI-Train) comprises 190 sentences, while the validation set (BIWI-Val) encompasses 24 sentences. The dataset is divided into two testing sets: BIWI-Test-A, which comprises 24 sentences articulated by six subjects observed during training, and BIWI-Test-B, which consists of 32 sentences

Table 1: Quantitative comparisons on the BIWI-Test-A and VOCA-Test datasets. The results of Lip-Vertex Error (LVE) and the upper-Face Dynamics Deviation (FDD) are reported. For both metrics, the lower the better.

Methods	BIWI-Test-A		VOCA-Test	
	LVE↓ $\times 10^{-4}$mm	FDD↓ $\times 10^{-5}$mm	LVE↓ $\times 10^{-5}$mm	FDD↓ $\times 10^{-7}$mm
VOCA [8]	6.5563	8.1816	4.9245	4.8447
MeshTalk [41]	5.9181	5.1025	4.5441	5.2062
FaceFormer [12]	5.3077	4.6408	4.1090	4.6675
CodeTalker [53]	4.7914	4.1170	3.9445	4.5422
SelfTalk [36]	4.2485	3.5761	3.2238	4.0912
KMTalk (Ours)	**3.9654**	**2.5446**	**2.2639**	**4.0594**

uttered by eight unseen subjects. *VOCASET* [8] consists of 480 paired audio-visual sequences from 12 subjects. Each sequence is recorded at a frame rate of 60fps and ranges in duration from 3 to 4 seconds. The 3D face mesh for each sequence consists of 5023 vertices. To ensure a fair comparison, we used identical training (VOCA-Train), validation (VOCA-Val), and testing (VOCA-Test) partitions as methods [12, 36, 53].

Baselines. We compare against current state-of-the-arts method, including VOCA [8], MeshTalk [41]. FaceFormer [12], CodeTalker [53], and SelfTalk [36]. Faceformer [12] employs a transformer-based model to incorporate long-term audio context and synthesizes sequential motions in an autoregressive manner. CodeTalker [53] introduces discrete motion priors to enable self-reconstruction of real facial movements, mitigating the issue of excessive smoothing in facial motion. SelfTalk [36] designs a learning-based recognizer to minimize the domain gap between diverse modalities.

Evaluation Metrics. Following CodeTalker [53] and SelfTalk [36], we adopt two metrics for the quantitative evaluation of speech-driven facial animation: *lip vertex error* (LVE) to measure lip synchronization and *upper-face dynamics deviation* (FDD) to assess the overall facial dynamics. The LVE for each frame is defined as the maximal L2 error among all lip vertices for each frame and takes the average over all frames. This L2 error is computed by comparing the predictions with the processed 3D face geometry data. FDD is introduced to quantify the variation in facial dynamics between a synthetic motion sequence and the reference sequence. The implementation of FDD is to calculate the difference between the variances of vertex offsets in the upper-face region and the variances of ground truth vertex offsets. In addition, we visualize the prediction results for qualitative evaluation.

Implementation Details. For a fair comparison, KMTalk operates at a frame rate of 30 fps on VOCASET and 25 fps on BIWI, following the setting of previous methods [12, 36, 53]. Also, it can naturally adapt to a higher frame rate, as shown in the supplementary materials. In the LKMA module, we first employ the Phoneme-based Localization method to process the raw audio and obtain key motion indices for data preprocessing, which costs less than 10 minutes on two datasets [8, 13]. Secondly, we train the Key Motion Decoder on a single NVIDIA

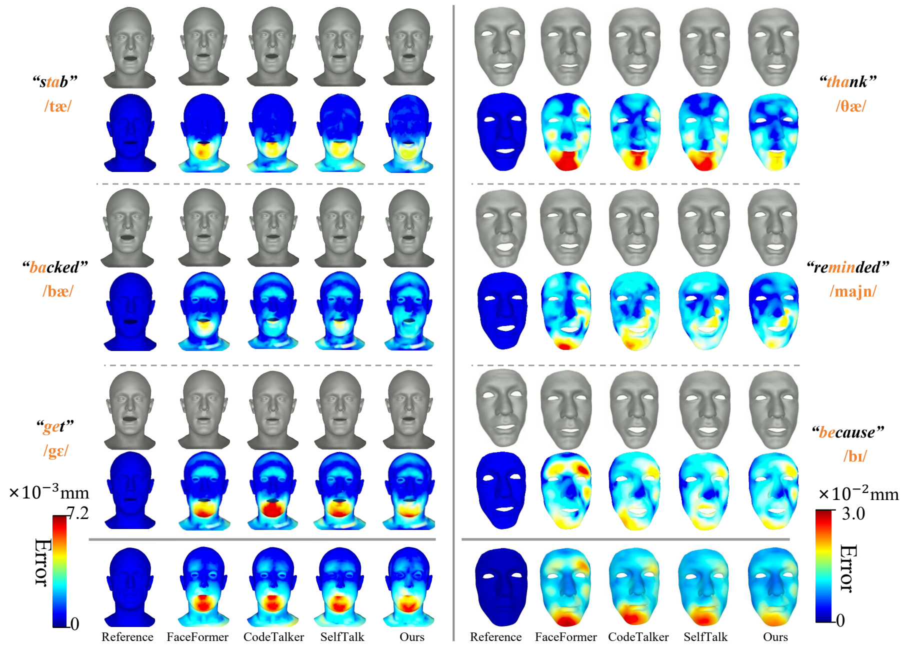


Fig. 3: Qualitative comparisons on VOCA-Test (left) and BIWI-Test-B (right). We provide visual comparisons of facial animations synchronized with six syllables extracted from the test speech sequences. The 1st, 3rd, and 5th rows display synthesized meshes and their corresponding ground-truths, while the 2nd, 4th, and 6th rows visualize the L2 loss for individual frames. Our method demonstrates more precise mouth movement on syllables like /æ/ that require a wide-open mouth. For syllables that start with a closed mouth and then slightly open, such as /bɪ/, our KMTalk generates more synchronized motion sequences visually. The last row visualizes the mean square errors of different methods across all sentences in the test set for a specific subject.

RTX 3090 for 200 epochs (about 2 hours) using the Adam optimizer [21]. The learning rate is initialized as 10^{-4}, and the mini-batch size is set to 1. In the CMC module, we train for 200 epochs (approximately 2 hours) with the same training settings as the Key Motion Decoder. It is noteworthy that, since the training of the two modules is independent of each other, we can train both modules concurrently to enhance training efficiency.

4.1 Comparisons against State-of-the-art Methods

Quantitative Comparisons. We computed the lip vertex error (LVE) and facial dynamics deviation (FDD) for all sequences within the BIWI-Test-A and VOCA-Test datasets. According to Table 1, our proposed KMTalk demonstrated lower errors compared to the alternative methods examined. Notably, the lip vertex error for our method on the VOCA-Test dataset is 30% lower than the recently introduced SelfTalk [36], and the FDD is 27% lower than SelfTalk on the BIWI-Test-A dataset, providing compelling evidence for the advantages of our proposed KMTalk. This indicates that our approach is more effective in achieving audio-visual alignment, thereby leading to improved lip synchronization.

Qualitative Comparisons. In Fig. 3, we visualize the output facial meshes from different methods as well as ground truths for reference. Additionally, we display error maps calculated from the vertex L2 loss between the generated and ground truth meshes. It is evident that our method consistently yields lower errors across different speech sequences, demonstrating its ability to generate more accurate facial animation sequences. Notably, for representative syllables (*e.g.* /æ/), KMTalk closely approximates ground truth, excelling in synthesizing accurate lip movements for syllables requiring significant mouth opening. Additionally, for syllables starting with mouth closure followed by a slight opening (e.g., /bɪ/), KMTalk produces more natural and synchronized motions. We recommend that readers watch the supplementary video for more detailed comparisons. It showcases KMTalk's capability to generate coherent, realistic animations with precise lip synchronization.

User Studies. A user study stands as a dependable evaluation method in the context of 3D talking faces. Following the strategy of Faceformer [12], we conduct pairwise comparisons between our method and baselines [12,36,53], as well as ground truths. This study encompassed the assessment of two key metrics: perceptual lip synchronization and facial realism. Participants were presented with side-by-side comparisons and were tasked with selecting the better facial animation based on their personal preferences. We computed the ratio of user preferences as a measurement of satisfaction evaluation on BIWI-Test-B and VOCA-Test. We randomly sampled 30 examples from each test set and compared the performance of KMTalk with four aforementioned settings on each sample. Therefore, we constructed a total of 240 different video pairs and randomly selected 24 video pairs for the two metrics assessments for each participant. Our user study involved 30 participants with a strong capability for audio-visual perception, resulting in 720 effective evaluation entries. As demonstrated in Table 2, our approach indicates superior perceptual lip synchronization and facial realism. For instance, a noteworthy 60.0% of users favored our lip synchronization method on BIWI-Test-B in comparison to SelfTalk [36]. Overall, it shows that KMTalk can generate more favorable facial animations from speeches.

4.2 Ablation Studies

In this section, we perform ablation studies on to evaluate the influence of different components within our proposed KMTalk framework on the quality of the generated 3D talking faces. The quantitative results on BIWI are in Table 3, and the qualitative results are in Fig. 4. In addition, the results of the ablation study on VOCA-Test can be found in the supplementary material. In Table 4, we further investigate the robustness of our approach to different Phoneme-based localization methods and the possible errors during phoneme extraction.

Table 2: User study results on BIWI-Test-B and VOCA-Test.

Method	Metric	BIWI-Test-B		VOCA-Test	
		competitor	ours	competitor	ours
Ours vs. FaceFormer	Lip Sync	24.4%	**75.6%**	26.7%	**73.3%**
	Realism	25.6%	**74.4%**	28.9%	**71.1%**
Ours vs. CodeTalker	Lip Sync	31.1%	**68.9%**	37.8%	**62.2%**
	Realism	27.8%	**72.2%**	35.6%	**64.4%**
Ours vs. SelfTalk	Lip Sync	40.0%	**60.0%**	43.3%	**56.7%**
	Realism	38.9%	**61.1%**	41.1%	**58.9%**
Ours vs. GT	Lip Sync	54.4%	**45.6%**	56.7%	**43.3%**
	Realism	52.2%	**47.8%**	56.7%	**43.3%**

Table 3: Ablation study for our components on BIWI-Test-A.

Phoneme-based Localization Method	**Key Motion-focused Decoder**	**Audio Guidance in CMC**	**LVE↓**	**FDD↓**
—	—	—	4.2485	3.5761
—	✓	✓	4.1648	2.8713
✓	—	✓	4.1381	2.9546
✓	✓	—	4.8859	3.2780
✓	✓	✓	**3.9654**	**2.5446**

What's the effect of the Phoneme-based localization method for key motion capture? The Phoneme-based Localization Method enables us to identify key frames of speech with notable facial expression transitions. We can replace it with uniform sampling, neglecting the crucial content information of the audio. Specifically, we experimented using uniform sampling at a rate of 33% to closely align with the number of key motions. For further comparisons under different numbers of sampled elements, please refer to the supplementary material. In Table 3, we observe degradation in all metrics with uniform sampling. This underscores the importance of accurate key motion capture in speech-driven talking face generation. Additionally, it showcases that the linguistic-based key motion capture is better equipped to mitigate audio-visual uncertainty and recover more precise facial motions.

What's the effect of the Key Motion Decoder? Our method employs a specialized key motion decoder to generate facial meshes based on keyframe indices obtained from phoneme-based localization methods. An alternative approach to generating key motions is to select corresponding facial meshes from a complete motion sequence produced by an existing method such as SelfTalk [36]. Table 3 demonstrates that both metrics deteriorated but still outperformed the state-of-the-art method SelfTalk. This indicates that the Key Motion Decoder can enhance the quality of key motion generation, resulting in more plausible facial animations. It also suggests that even if the captured key motions are not accurate enough, the CMC module can further refine output full motions.

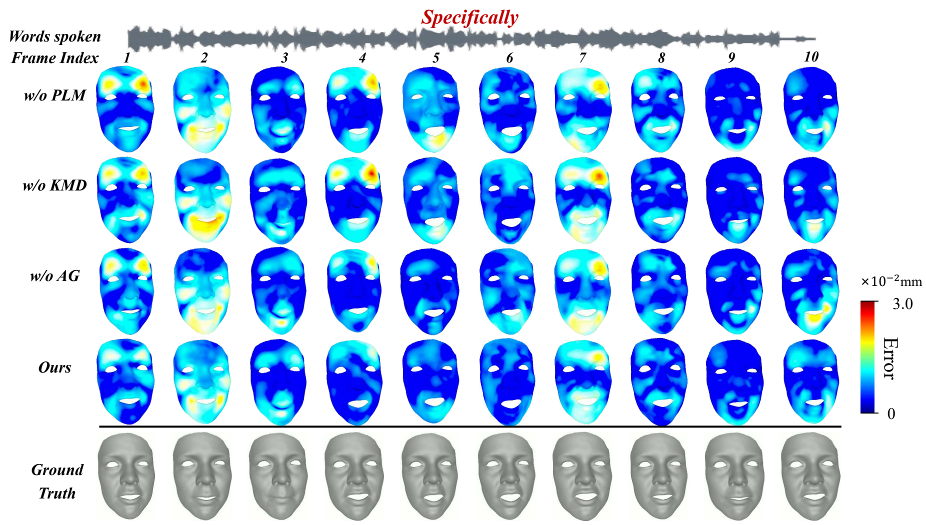


Fig. 4: Qualitative ablation studies on the input speech "specifically". For each method variant, we removed one of three modules: PLM (Phoneme-based Localization Method), KMD (Key Motion-focused Decoder), and AG (Audio Guidance in CMC). Error maps between generated and the ground-truth mesh sequence were visualized. Our final model yielded the best results, showcasing the effectiveness of each module.

What's the effect of audio guidance in the Cross-modal Motion Completion? The Cross-modal Motion Completion leverages audio features to guide the completion of the full motion sequence. To assess its usefulness, we implement a method variant that removes the audio feature guidance. Table 3 demonstrates a notable degradation in both metrics, particularly with a 24% increase in Lip Vertex Error (LVE) and a 27% increase in upper-face dynamics deviation (FDD). This suggests that audio information plays a crucial role in refining fine-grained lip movements and enhancing audio-visual consistency and temporal smoothness. Despite the degradation, the FDD metric can still outperform state-of-the-art methods, underscoring the significance of key motion capture for achieving temporally coherent full motion synthesis.

Is our method sensitive to different Phomene-based Localization methods? We experimented with different Automatic Speech Recognition (ASR) models, such as Auto-avsr [28] and Whisper [39]. The results in the first three rows of Table 4 show that our KMTalk method consistently maintained high performance, achieving at least a 21% improvement in FDD regardless of the ASR model used [28, 39]. This highlights the robustness of our approach across various ASR models. Besides, to simulate phoneme localization deviations, we shifted all key motion indices extracted by Auto-avsr [28] one frame to the right. The results in the last row of Table 4 indicate negligible variations. This demonstrates the robustness of our method to inaccurate key frame localization.

Table 4: Robust analysis of Phoneme-based Localization on BIWI-Test-A.

Methods	LVE↓ $\times 10^{-4}$mm	FDD↓ $\times 10^{-5}$mm
Auto-avsr [28]	**3.9654**	**2.5446**
Whisper-large [39]	4.0718	2.8141
Whisper-tiny [39]	4.0643	2.8083
Auto-avsr+offset	3.9991	2.6420

Table 5: The results of integrating our proposed KMTalk with existing methods on BIWI-Test-A.

Methods		LVE↓ $\times 10^{-4}$mm	FDD↓ $\times 10^{-5}$mm
FaceFormer [12]	Original	5.3077	4.6408
	After	**5.2793**	**4.2654**
CodeTalker [53]	Original	4.7914	4.1170
	After	**4.5096**	**3.9043**
SelfTalk [36]	Original	4.2485	3.5761
	After	**4.1122**	**2.8668**

4.3 Integration with Existing Methods

Existing approaches focus on enhancing prediction outcomes by designing elaborate priors or the learning-based recognizer, which may be highly coupled with the proposed architecture of these methods. Our KMTalk introduces a new learning strategy of speech-driven talking face generation, which is orthogonal to these approaches. Therefore, we can explore whether performance can be enhanced by applying our progressive learning scheme without the need for additional fine-tuning of their models. Detailed implementation is in the supplementary material. As shown in Table 5, the results of existing methods are improved after integration with our proposed progressive learning mechanism utilizing key motion embeddings. This further emphasizes the efficacy of our design.

5 Conclusion

In this work, we introduce KMTalk, a novel method for progressively learning 3D facial animation from speeches using key motion embeddings. It incorporates linguistic priors for key motion generation and extends them to a full motion sequence via data-driven priors. We propose phoneme-based localization methods to determine the temporal position of key facial motions, improving lip-speech synchronization by aligning motion transitions with phoneme changes. Additionally, we design a cross-modal facial motion completion module that synthesizes the entire motion sequence from key motions and audio features, enhancing lip-speech synchronization and motion coherence. Extensive evaluations of the datasets demonstrate KMTalk's superiority over existing methods, producing more accurate and realistic animations. Moreover, coupling our idea with existing methods consistently improves performance, further verifying the efficacy of our proposed progressive learning mechanism based on key motion acquisition. Although the proposed method has demonstrated its robustness to inaccurate keyframe localization, it may encounter errors in dialect variations. Integrating advanced ASR(Automatic Speech Recognition) technology in the future could enhance its adaptability to various speech patterns.

Acknowledgement The research is partially supported by National Key Research and Development Program of China (2023YFC3502900), National Natural Science Foundation of China (No.62176093, 61673182), Key Realm Research and Development Program of Guangzhou (No.202206030001), Guangdong-Hong Kong-Macao Joint Innovation Project (No.2023A0505030016), Guangdong Natural Science Foundation (No. 2024A1515012217).

References

1. Alghamdi, M.M., Wang, H., Bulpitt, A.J., Hogg, D.C.: Talking head from speech audio using a pre-trained image generator. In: ACM MM. pp. 5228–5236 (2022)
2. Baevski, A., Zhou, Y., Mohamed, A., Auli, M.: wav2vec 2.0: A framework for self-supervised learning of speech representations. Advances in neural information processing systems **33**, 12449–12460 (2020)
3. Bao, L., Zhang, H., Qian, Y., Xue, T., Chen, C., Zhe, X., Kang, D.: Learning audio-driven viseme dynamics for 3d face animation. arXiv preprint arXiv:2301.06059 (2023)
4. Cao, Y., Tien, W.C., Faloutsos, P., Pighin, F.: Expressive speech-driven facial animation. ACM Trans. Graph **24**(4), 1283–1302 (2005)
5. Chen, L., Cui, G., Liu, C., Li, Z., Kou, Z., Xu, Y., Xu, C.: Talking-head generation with rhythmic head motion. In: ECCV. pp. 35–51. Springer (2020)
6. Chung, J.S., Zisserman, A.: Out of time: automated lip sync in the wild. In: ACCV. pp. 251–263. Springer (2017)
7. Cohen, M.M., Clark, R., Massaro, D.W.: Animated speech: Research progress and applications. In: AVSP (2001)
8. Cudeiro, D., Bolkart, T., Laidlaw, C., Ranjan, A., Black, M.J.: Capture, learning, and synthesis of 3d speaking styles. In: CVPR. pp. 10101–10111 (2019)
9. Dai, G., Zhang, Y., Wang, Q., Du, Q., Yu, Z., Liu, Z., Huang, S.: Disentangling writer and character styles for handwriting generation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5977–5986 (2023)
10. Das, D., Biswas, S., Sinha, S., Bhowmick, B.: Speech-driven facial animation using cascaded gans for learning of motion and texture. In: ECCV. pp. 408–424. Springer (2020)
11. Edwards, P., Landreth, C., Fiume, E., Singh, K.: Jali: an animator-centric viseme model for expressive lip synchronization. ACM Trans. Graph **35**(4), 1–11 (2016)
12. Fan, Y., Lin, Z., Saito, J., Wang, W., Komura, T.: Faceformer: Speech-driven 3d facial animation with transformers. In: CVPR. pp. 18770–18780 (2022)
13. Fanelli, G., Gall, J., Romsdorfer, H., Weise, T., Van Gool, L.: A 3-d audio-visual corpus of affective communication. IEEE Transactions on Multimedia **12**(6), 591–598 (2010)
14. Fisher, C.G.: Confusions among visually perceived consonants. Journal of speech and hearing research **11**(4), 796–804 (1968)
15. Guo, Y., Chen, K., Liang, S., Liu, Y.J., Bao, H., Zhang, J.: Ad-nerf: Audio driven neural radiance fields for talking head synthesis. In: ICCV. pp. 5784–5794 (2021)
16. Habibie, I., Xu, W., Mehta, D., Liu, L., Seidel, H.P., Pons-Moll, G., Elgharib, M., Theobalt, C.: Learning speech-driven 3d conversational gestures from video. In: Proceedings of the 21st ACM International Conference on Intelligent Virtual Agents. pp. 101–108 (2021)

17. Iverson, P., Kuhl, P.K.: Perceptual magnet and phoneme boundary effects in speech perception: Do they arise from a common mechanism? Perception & psychophysics **62**, 874–886 (2000)
18. Ji, X., Zhou, H., Wang, K., Wu, Q., Wu, W., Xu, F., Cao, X.: Eamm: One-shot emotional talking face via audio-based emotion-aware motion model. In: ACM SIGGRAPH. pp. 1–10 (2022)
19. Ji, X., Zhou, H., Wang, K., Wu, W., Loy, C.C., Cao, X., Xu, F.: Audio-driven emotional video portraits. In: CVPR. pp. 14080–14089 (2021)
20. Karras, T., Aila, T., Laine, S., Herva, A., Lehtinen, J.: Audio-driven facial animation by joint end-to-end learning of pose and emotion. ACM Trans. Graph **36**(4), 1–12 (2017)
21. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
22. Lewis, J.: Automated lip-sync: Background and techniques. IEEE Trans Vis Comput Graph **2**(4), 118–122 (1991)
23. Liang, B., Pan, Y., Guo, Z., Zhou, H., Hong, Z., Han, X., Han, J., Liu, J., Ding, E., Wang, J.: Expressive talking head generation with granular audio-visual control. In: CVPR. pp. 3387–3396 (2022)
24. Liu, C.: An analysis of the current and future state of 3d facial animation techniques and systems (2009)
25. Liu, X., Xu, Y., Wu, Q., Zhou, H., Wu, W., Zhou, B.: Semantic-aware implicit neural audio-driven video portrait generation. In: ECCV. pp. 106–125. Springer (2022)
26. Liu, Y., Xu, F., Chai, J., Tong, X., Wang, L., Huo, Q.: Video-audio driven real-time facial animation. ACM Trans. Graph **34**(6), 1–10 (2015)
27. Lu, L., Wu, R., Lin, H., Lu, J., Jia, J.: Video frame interpolation with transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3532–3542 (2022)
28. Ma, P., Haliassos, A., Fernandez-Lopez, A., Chen, H., Petridis, S., Pantic, M.: Auto-avsr: Audio-visual speech recognition with automatic labels. arxiv 2023. arXiv preprint arXiv:2303.14307
29. Massaro, D., Cohen, M., Tabain, M., Beskow, J., Clark, R.: 12 animated speech: research progress and applications (2012)
30. Mattheyses, W., Verhelst, W.: Audiovisual speech synthesis: An overview of the state-of-the-art. Speech Communication **66**, 182–217 (2015)
31. McAuliffe, M., Sonderegger, M.: English mfa acoustic model v2.2.1. Tech. rep., https://mfa-models.readthedocs.io/acoustic/English/EnglishMFAacousticmodelv2_2_1.html (May 2023)
32. Mo, C.A., Hu, K., Long, C., Wang, Z.: Continuous intermediate token learning with implicit motion manifold for keyframe based motion interpolation. In: CVPR. pp. 13894–13903 (2023)
33. Niklaus, S., Mai, L., Liu, F.: Video frame interpolation via adaptive convolution. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 670–679 (2017)
34. Nocentini, F., Ferrari, C., Berretti, S.: Learning landmarks motion from speech for speaker-agnostic 3d talking heads generation. arXiv preprint arXiv:2306.01415 (2023)
35. Pang, Y., Zhang, Y., Quan, W., Fan, Y., Cun, X., Shan, Y., Yan, D.m.: Dpe: Disentanglement of pose and expression for general video portrait editing. In: CVPR. pp. 427–436 (2023)

36. Peng, Z., Luo, Y., Shi, Y., Xu, H., Zhu, X., Liu, H., He, J., Fan, Z.: Selftalk: A self-supervised commutative training diagram to comprehend 3d talking faces. arXiv preprint arXiv:2306.10799 (2023)
37. Pham, H.X., Wang, Y., Pavlovic, V.: End-to-end learning for 3d facial animation from speech. In: ICMI. pp. 361–365 (2018)
38. Prajwal, K., Mukhopadhyay, R., Namboodiri, V.P., Jawahar, C.: A lip sync expert is all you need for speech to lip generation in the wild. In: ACM MM. pp. 484–492 (2020)
39. Radford, A., Kim, J.W., Xu, T., Brockman, G., McLeavey, C., Sutskever, I.: Robust speech recognition via large-scale weak supervision. In: International Conference on Machine Learning. pp. 28492–28518. PMLR (2023)
40. Richard, A., Lea, C., Ma, S., Gall, J., De la Torre, F., Sheikh, Y.: Audio-and gaze-driven facial animation of codec avatars. In: WACV. pp. 41–50 (2021)
41. Richard, A., Zollhöfer, M., Wen, Y., De la Torre, F., Sheikh, Y.: Meshtalk: 3d face animation from speech using cross-modality disentanglement. In: ICCV. pp. 1173–1182 (2021)
42. Shen, S., Li, W., Zhu, Z., Duan, Y., Zhou, J., Lu, J.: Learning dynamic facial radiance fields for few-shot talking head synthesis. In: ECCV. pp. 666–682. Springer (2022)
43. Tanaka, H., Nakamura, S., et al.: The acceptability of virtual characters as social skills trainers: usability study. JMIR human factors **9**(1), e35358 (2022)
44. Tang, J., Dai, A., Nie, Y., Markhasin, L., Thies, J., Nießner, M.: Dphms: Diffusion parametric head models for depth-based tracking. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1111–1122 (2024)
45. Taylor, S., Kim, T., Yue, Y., Mahler, M., Krahe, J., Rodriguez, A.G., Hodgins, J., Matthews, I.: A deep learning approach for generalized speech animation. ACM Trans. Graph **36**(4), 1–11 (2017)
46. Taylor, S.L., Mahler, M., Theobald, B.J., Matthews, I.: Dynamic units of visual speech. In: ACM SIGGRAPH. pp. 275–284 (2012)
47. Vougioukas, K., Petridis, S., Pantic, M.: Realistic speech-driven facial animation with gans. International Journal of Computer Vision **128**, 1398–1413 (2020)
48. Wang, J., Qian, X., Zhang, M., Tan, R.T., Li, H.: Seeing what you said: Talking face generation guided by a lip reading expert. In: CVPR. pp. 14653–14662 (2023)
49. Wang, Q., Fan, Z., Xia, S.: 3d-talkemo: Learning to synthesize 3d emotional talking head. arXiv preprint arXiv:2104.12051 (2021)
50. Wang, S., Li, L., Ding, Y., Yu, X.: One-shot talking face generation from single-speaker audio-visual correlation learning. In: AAAI. vol. 36, pp. 2531–2539 (2022)
51. Wen, S., Liu, W., Yang, Y., Huang, T., Zeng, Z.: Generating realistic videos from keyframes with concatenated gans. IEEE Transactions on Circuits and Systems for Video Technology **29**(8), 2337–2348 (2018)
52. Wu, H., Zhou, S., Jia, J., Xing, J., Wen, Q., Wen, X.: Speech-driven 3d face animation with composite and regional facial movements. arXiv preprint arXiv:2308.05428 (2023)
53. Xing, J., Xia, M., Zhang, Y., Cun, X., Wang, J., Wong, T.T.: Codetalker: Speech-driven 3d facial animation with discrete motion prior. In: CVPR. pp. 12780–12790 (2023)
54. Xu, Y., Feng, A.W., Marsella, S., Shapiro, A.: A practical and configurable lip sync method for games. In: Proceedings of Motion on Games, pp. 131–140 (2013)

55. Yi, R., Ye, Z., Zhang, J., Bao, H., Liu, Y.J.: Audio-driven talking face video generation with learning-based personalized head pose. arXiv preprint arXiv:2002.10137 (2020)
56. Yin, S., Wu, C., Yang, H., Wang, J., Wang, X., Ni, M., Yang, Z., Li, L., Liu, S., Yang, F., et al.: Nuwa-xl: Diffusion over diffusion for extremely long video generation. arXiv preprint arXiv:2303.12346 (2023)
57. Yu, D., Li, X., Zhang, C., Liu, T., Han, J., Liu, J., Ding, E.: Towards accurate scene text recognition with semantic reasoning networks. In: CVPR. pp. 12113–12122 (2020)
58. Yue, X., Kuang, Z., Lin, C., Sun, H., Zhang, W.: Robustscanner: Dynamically enhancing positional clues for robust text recognition. In: ECCV. pp. 135–151. Springer (2020)
59. Zhang, B., Qi, C., Zhang, P., Zhang, B., Wu, H., Chen, D., Chen, Q., Wang, Y., Wen, F.: Metaportrait: Identity-preserving talking head generation with fast personalized adaptation. In: CVPR. pp. 22096–22105 (2023)
60. Zhang, C., Ni, S., Fan, Z., Li, H., Zeng, M., Budagavi, M., Guo, X.: 3d talking face with personalized pose dynamics. IEEE Trans Vis Comput Graph. (2021)
61. Zhou, H., Sun, Y., Wu, W., Loy, C.C., Wang, X., Liu, Z.: Pose-controllable talking face generation by implicitly modularized audio-visual representation. In: CVPR. pp. 4176–4186 (2021)