ProsodyTalker: 3D Visual Speech Animation via Prosody Decomposition

Zonglin Li, Xiaoqian Lv, Qinglin Liu, Quanling Meng, Xin Sun*, Shengping Zhang

Harbin Institute of Technology
{zonglin.li, xiaoqian.lv}@hit.edu.cn, qinglin.liu@outlook.com,
{quanling.meng, sunxintyc, s.zhang}@hit.edu.cn

Abstract

Most existing 3D visual speech animation methods synthesize lip movements synchronized with speech, which however neglect head poses and therefore degrade the animation realism. The animation of head poses presents two primary challenges: (1) the intricate mapping between speech and head poses remains poorly understood and (2) the absence of 4D face datasets featuring realistic head poses. Inspired by prosody decomposition in speech processing, we discern that head movements correlate with the fundamental frequency (F0) of speech prosody, while lip movements align with the language content. These observations motivate us to propose a novel framework, dubbed ProsodyTalker, that concurrently synthesizes lip and head movements, grounded in the principles of prosody decomposition. The core idea is first to adopt information perturbation to explicitly decompose the speech prosody into pose-related F0 and lip-related language content. Then, an autoregressive content-oriented fusion decoder is employed to enhance lip synchronization in the synthesized facial sequences. To synthesize head poses, we design a transformer-based variational autoencoder to learn a latent distribution of facial sequences and propose an F0-conditioned latent diffusion model to establish a probabilistic mapping from F0 to pose-related latent codes. Furthermore, we contribute a large-scale 4D face dataset containing bunches of variations in identities, head poses, and facial motions. Extensive experiments show that our method achieves more realistic animation than state-of-the-art methods.

Introduction

3D visual speech animation has experienced notable growth in academia and industry, owing to its diverse applications in virtual reality, film, and game production. Several methods (Cudeiro et al. 2019; Fan et al. 2022; Peng et al. 2023a) have shown promising performance in lip synchronization, whereas they neglect head pose synthesis and therefore degrade the animation realism in real-world scenarios. As for head pose synthesis, there are two main challenges: (1) speech is an implicit control signal for head poses (Henter, Alexanderson, and Beskow 2020; Zhang et al. 2021a), indicating that directly learning mappings on such signals inherently introduces uncertainty in head pose synthesis. (2) since existing 4D face datasets have no head poses and real audiovisual animation data is challenging to collect, the data scarcity problem is intrinsically inevitable.

Our initial analysis focuses on discerning specific signals that correlate with lip movements and head poses within speech. In cognitive science and psycholinguistics research (Yehia, Kuratate, and Vatikiotis-Bateson 2002; Graf et al. 2002), it has been demonstrated that lip movements are closely tied to content semantics, whereas head poses are more intricately linked to speech intonation. For instance, emphasis on a word is frequently accompanied by a nod of the head, and a rising voice at the end of a phrase is often underlined with an upward head movement, analogous to variations in F0 within speech prosody. Drawing inspiration from prosody decomposition in speech processing, we conduct a statistical analysis to demonstrate how lip movements and head pose correlate with the language content and F0. These findings support the integration of language content and F0 in 3D visual speech animation.

Given these observations, we propose a novel framework, dubbed ProsodyTalker, which simultaneously leverages language content and F0 to synthesize co-speech lip movements and head poses. We begin by applying information perturbation to distinctly decompose speech prosody into pose-related F0 and lip-related language content. Given the correlation between lip movements and language content, we combine multi-head attention with inductive biases to align content-lip modalities and integrate temporal motion context for enhanced lip synchronization. Meanwhile, we leverage the strengths of latent space-based and conditional diffusion-based methods and propose an F0-conditioned latent diffusion model to learn a probabilistic mapping from F0 to pose-related latent codes. Different from GeneFace++ (Ye et al. 2023), the continuous nature of F0 embedding enables a more accurate representation of F0 variations, resulting in more expressive head pose synthesis, with temporal smoothness ensured by the latent diffusion model.

Existing 3D visual speech animation approaches continue to rely on 4D face datasets for model training. However, capturing high-quality audiovisual data remains notably time-intensive and resource-demanding. Existing 4D face datasets (e.g., VOCASET (Cudeiro et al. 2019) and BIWI (Fanelli et al. 2010)) generally lack head pose variations, making it hard to train the models to capture plausible

*Corresponding author (sunxintyc@hit.edu.cn)

head poses. Concurrently, 2D talking face datasets such as MEAD (Wang et al. 2020) and HDTF (Zhang et al. 2021b) only provide the audio and video files, without incorporating 3D facial parameters. Due to the scarcity of high-quality audiovisual data, we construct a large-scale 4D face dataset with diverse identities, head poses, and facial motions.

- We discern that lip movements are correlated to language content and head poses are associated with F0 variations. Upon these observations, we explicitly split visual speech animation into content-oriented lip synchronization and F0-conditioned head pose synthesis.
- We propose ProsodyTalker, a novel framework that jointly synthesizes precise lip synchronization with diverse head pose sequences from speech. To the best of our knowledge, we are the first to incorporate F0 into 3D visual speech animation tasks.
- We design a transformer-based variational autoencoder to learn a low-dimensional latent distribution of diverse facial sequences. Then, we propose an F0-conditioned latent diffusion model to establish a probabilistic mapping from F0 to pose-related latent codes.
- We construct a large-scale 4D face dataset from 2D talking head videos with diverse identities, head poses, and facial motions. This dataset also provides a referable solution to alleviate the data scarcity problem of similar 3D video synthesis tasks.

Related Work

Visual Speech Animation. Most existing visual speech animation approaches focus on learning a mapping between speech and facial mesh to achieve lip synchronization. VOCA (Cudeiro et al. 2019) leverages temporal convolutions to synthesize realistic lip movements from speeches. MeshTalk (Richard et al. 2021) considers longer audio contexts and trains a categorical latent space to disentangle audio-correlated and audio-uncorrelated information. FaceFormer (Fan et al. 2022) captures the long-term speech context based on a transformer and then employs an autoregressive manner to synthesize lip movements. CodeTalker (Xing et al. 2023) introduces a discrete motion prior codebook and synthesizes lip movements based on a code query strategy. EmoTalk (Peng et al. 2023b) leverages an emotion-disentanglement module to separate speech content and emotion, producing high-quality emotional blendshape coefficients. SelfTalk (Peng et al. 2023a) synthesizes coherent and visually comprehensible lip movements by reducing the domain gap between different modalities. However, these approaches drive facial animation solely with speech features, struggling to directly correspond with head poses.

Motion Generative Models. Early motion generation work (Aksan, Kaufmann, and Hilliges 2019; Mao et al. 2019) usually adopts deterministic motion modeling to synthesize a single motion, which suffers from mode averaging problems. To address these problems, recent efforts have shifted towards deep generative models. MotionDiffuse (Zhang et al. 2022) is the first conditional motion diffusion model to achieve motion synthesis with an arbitrary length depending on the motion duration. MDM (Tevet et al. 2023) proposes a motion diffusion model on raw motion data to learn the correlations between motion and input conditions. However, these diffusion models do not apply to raw motion data with potential noise, making them susceptible to being misled by outliers. In addition, directly applying the diffusion model to the raw motion data suffers from high computational overheads and low inference speed.

Motivation and Statistical Analysis

Given that synchronized lip movements and head poses consistently accompany speech, it is rational to pursue a transformation approach that predicts these motions from speech. To advance this approach, we introduce a novel perspective by incorporating decomposed prosody elements from speech into the synthesis of lip movements and head poses.

To support our motivation, two principal questions must be elucidated: (1) how to quantify the offsets of a specific location in facial motions and (2) what is the correlation between decomposed prosody elements and facial motions. In this section, we endeavor to illuminate these questions over 400 sequences from the collected dataset. We first define a facial motion metric to distinguish the vertex offset variations within the facial sequence. Given a reference facial sequence $\boldsymbol{F} \in R^{L\times V\times 3}$ with L frames and a neutral face $\boldsymbol{h} \in R^{V\times 3}$ in the form of 3D coordinates of V vertices, we define the vertex offset $\boldsymbol{D} \in R^{L\times V}$ as metric to represent the facial motions against the neutral face geometry:

$$\boldsymbol{D}(l, v) = \|\boldsymbol{F}(l, v, :) - \boldsymbol{h}(v, :)\|_2^2. \tag{1}$$

This process records the spatiotemporal deviations of vertices from their neutral geometry, representing facial motion dynamics. After analyzing multiple vertices, we select a stable vertex located centrally within the lip and forehead regions to capture lip movements and head poses, respectively.

Upon the given speech, we adopt information perturbation to explicitly decompose the speech prosody into F0 and language content, which is explained in the Methodology. As demonstrated in DMRN (Tian et al. 2018), synchronization of variations across both audio and visual modalities can effectively demonstrate cross-modality correspondence. To unveil the correlation between decomposed prosody elements and facial motions, we compute their averaged offset variations in the corresponding periods (*the offset variation is illustrated in the Technical Appendix*). For decomposed prosody elements, we separately divide them into k periods at the same interval w_b and calculate the amplitudes $\boldsymbol{X} \in R^{n\times k}$ across different periods, where n denotes the number of audio clips. The averaged offset variation h_i for $i \in \{1, \cdots, k\}$ is calculated as:

$$h_i = \frac{\sum_{j=1}^{n} \boldsymbol{X}(j, i)}{n}. \tag{2}$$

For facial motions, we compute the offsets of the selected vertices in the lip and forehead regions via Equation 1 and similarly calculate the average offset variations for language content and F0. As shown in Figures 1(a) and 1(b), lip motion variations correlate positively with language content

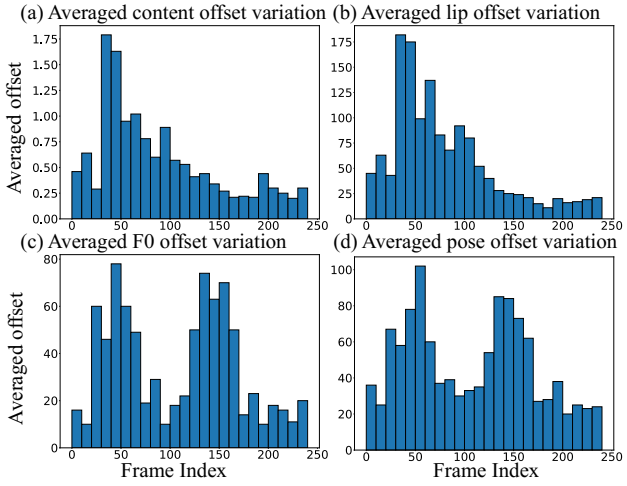

Figure 1: Averaged offset variations in (a) language content, (b) lip movement, (c) F0, and (d) head pose. we observe a positive correlation between content variations and lip movements, as well as F0 variations and head poses.

variations, indicating a strong relationship. Meanwhile, Figures 1(c) and 1(d) demonstrate a positive correlation between head poses and F0 changes. To further quantify these correlations, we calculate the Pearson correlation coefficient between the two pairs of variables and obtain a score of 0.747 between the language content and lip movements and 0.632 between F0 and head poses. This result highlights the interdependence of these variables and confirms that F0 is a reliable indicator of head poses during speech.

Methodology

Overview

As illustrated in Figure 2(a), given a neutral facial mesh $\boldsymbol{h}$ and an input speech $\mathcal{X}$, our goal is to synthesize animations $\tilde{\boldsymbol{F}}_{1:L}$ of L frames that contain well-synchronized lip movements and head poses, where each frame $\tilde{\boldsymbol{f}}_l \in R^{V\times 3}$ denotes the 3D offsets of V vertices over $\boldsymbol{h}$. In Figure 2(b), the prosody decomposition encoder contains three parts: content, style, and F0 encoder. The content encoder extract content embedding $\boldsymbol{C}_{1:L}$, where each $\boldsymbol{c}_l \in R^{d_c}$. The style and F0 encoders separately extract time-invariant style embedding $\boldsymbol{s} \in R^{d_s}$ and F0 embedding $\boldsymbol{y} \in R^{d_f}$. In Figure 2(c), the content-oriented fusion decoder predicts zero-posed facial sequences $\hat{\boldsymbol{F}}_{1:L}$ based on $\boldsymbol{C}_{1:L}$ and $\boldsymbol{s}$ via an autoregressive manner. In Figure 2(d), the F0-conditioned latent diffusion module synthesizes animation sequences with head poses $\tilde{\boldsymbol{F}}_{1:L}$ conditioned on the F0 embedding $\boldsymbol{y}$.

Prosody Decomposition Encoder

Content Encoding. The content encoder aims to remove speaker-related information and extract relatively pure content-related features. To this end, we follow NANSY (Choi et al. 2021) and employ information perturbation to remove unwanted information by signal processing beforehand. Specifically, we introduce three signal processing functions: pitch randomization (*pr*), formant shifting (*fs*), and random frequency shaping using a parametric equalizer (*peq*). Since speech signals exhibit unpredictable pitch fluctuations due to different speakers, the pitch randomization function is employed to expose the encoder to a diverse range of pitch patterns, which enhances the ability to filter out speaker-related information. Additionally, by shifting formants dynamically, the encoder can better accommodate diverse speech contexts, enhancing its generalization across different recording environments. Random frequency shaping acts as a form of regularization, preventing the model from relying too heavily on specific frequency components that might not be generalizable. In summary, the information perturbation can be described as:

$$\hat{\mathcal{X}} = pr(fs(peq(\mathcal{X})), \tag{3}$$

where $\hat{\mathcal{X}}$ is the perturbed waveform that is perceived as speaker-irrelevant and content-preserved.

The content encoder is a fully convolution network that encodes $\hat{\mathcal{X}}$ into a compact content representation. As shown in Figure 2(b), $\hat{\mathcal{X}}$ first goes through a convolution layer and then is fed into residual blocks with 2-stride max-pooling layers for downsampling. The architecture of the residual block incorporates consistent weight scaling across layers to enhance generalization, where LeakyReLU activation and weight normalization are applied to all convolution layers.

F0 Encoding. The F0 encoder aims to capture temporal variations in F0 over time. Due to the irregular periodicity of the glottal pulse, artifacts like jitter often appear in speech, typically manifesting irregularities in the F0 contour. To mitigate this effect, we compute the mean and variance of F0 across the constructed dataset, which serves as the normalization coefficient for the input speech. Since speech signals exhibit inherent variability due to various speakers, we incorporate the speaker style into F0 encoding to adapt to this variability. As shown in Figure 2(a), the input speech $\mathcal{X}$ is first transformed into a mel spectrogram. Then, a style encoder is employed to extract the style embedding from the mel spectrogram. In summary, the normalized speech is fed into the F0 encoder, and the F0 embedding is obtained based on the style embedding. As depicted in Figure 2(b), the F0 encoder consists of two encoding blocks, each of which is composed of a convolution layer with a PReLU activation layer and a channel normalization layer for style embedding injection. Following Expressive-vc (Ning et al. 2023), we compute the feature matching and multi-resolution short-time Fourier transform loss for the encoder optimization.

Content-Oriented Fusion Decoder

Motion Context Encoding. Inspired by FaceFormer (Fan et al. 2022), we incorporate motion context encoding (MCE) as temporal information to improve the generalization of the proposed method to longer sequence synthesis. Specifically, we first project the past facial motion $\hat{\boldsymbol{f}}_{t-1}$ into a d-dimensional latent code via a motion encoder that is an MLP block. Then, to autoregressively synthesize lip movements,

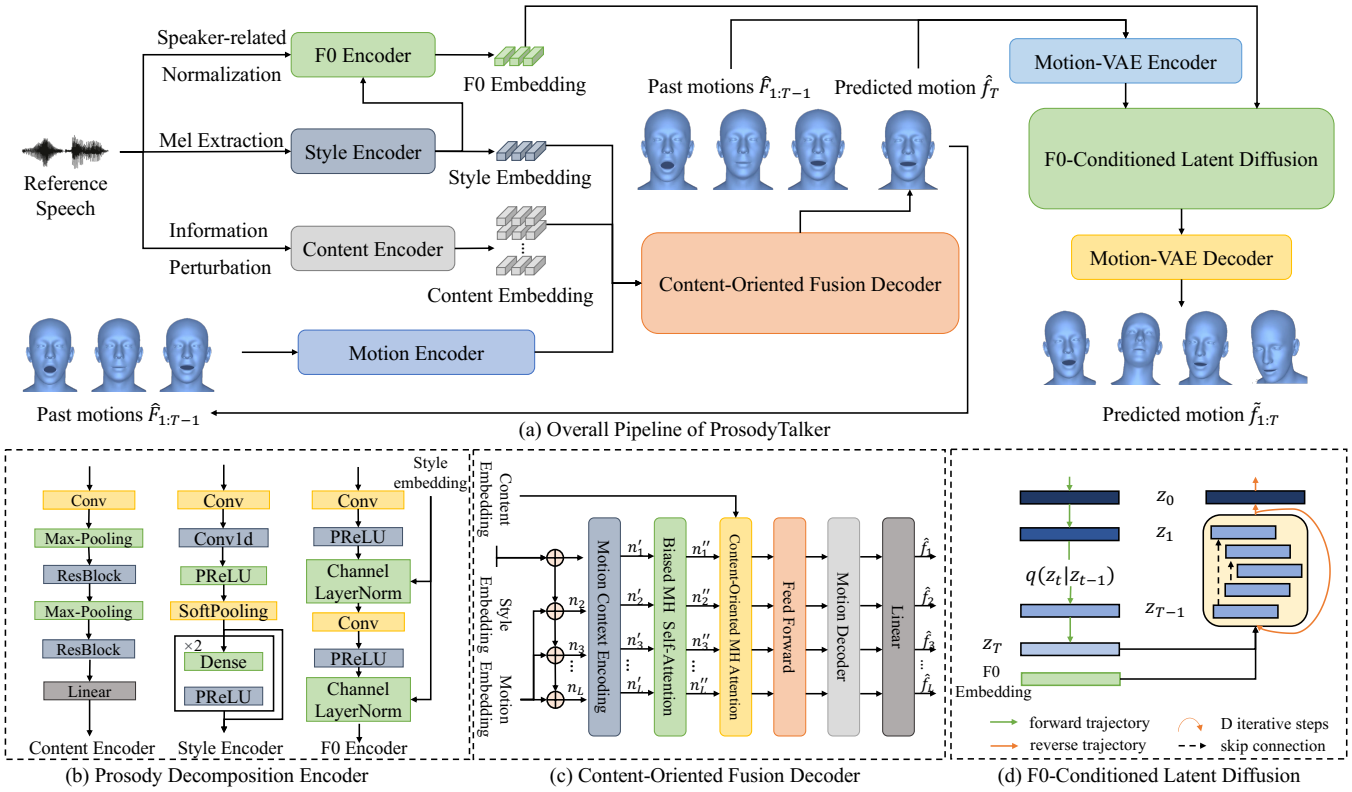

Figure 2: Overall pipeline of ProsodyTalker. Given an input speech, we propose a prosody decomposition encoder that separately obtains content, style, and F0 embedding. The content-oriented fusion decoder enhances lip synchronization in the synthesized facial sequences. The F0-conditioned latent diffusion model is designed to further synthesize head poses.

we concatenate the content embedding $\boldsymbol{c}_t$ and style embedding $\boldsymbol{s}$ to a d-dimensional vector and fed it into the current motion embedding $\boldsymbol{n}_t$:

$$\boldsymbol{n}_t = \begin{cases} \left(\boldsymbol{W}^n \hat{\boldsymbol{f}}_{t-1} + \boldsymbol{e}^n\right) + \mathcal{C}(\boldsymbol{c}_t, \boldsymbol{s}) & 1 < t \leq L \\ \mathcal{C}(\boldsymbol{c}_t, \boldsymbol{s}) & t = 1 \end{cases}, \quad (4)$$

where $\boldsymbol{W}^n$ is a weight matrix, $\boldsymbol{e}^n$ denotes a bias, $\mathcal{C}(\cdot,\cdot)$ indicates feature concatenation. Given the limited generalization capability of transformer-based approaches for longer sequences (Press, Smith, and Lewis 2021), we incorporate the motion context into the latent space to enhance generalization. To this end, we fuse explicit temporal information into the model via MCE, resulting in a timestamp vector as:

$$MCE_{(t,i)} = \frac{1}{\sqrt{d}} \left(\boldsymbol{n}_t^i \boldsymbol{W}^q\right) \left(\boldsymbol{n}_{t-1}^i \boldsymbol{W}^k\right)^T + \hat{b}, \quad (5)$$

where $\boldsymbol{W}^q$ and $\boldsymbol{W}^k$ denote the weight matrices, i represents the dimension index and $\hat{b}$ stands for a bias item. This vector contributes to improved model generalization for lip movement synthesis and enhances the handling of varying sequence lengths. Instead of assigning a distinct positional identifier to each token, the proposed MCE infuses motion context periodically into the encoded vector:

$$\boldsymbol{n}_t' = \boldsymbol{n}_t + MCE(t). \quad (6)$$

Dynamic Contextual Attention. Inspired by ALiBi (Press, Smith, and Lewis 2021), a multi-head self-attention mechanism with temporal bias is employed to integrate motion context encoding into multi-head attention layers, assigning higher weights to closer information and benefiting motion modeling. Upon the previous encoded motion context $\boldsymbol{N}'_{1:t} = (\boldsymbol{n}'_1, \ldots, \boldsymbol{n}'_t)$, we first linearly projects $\boldsymbol{N}'_{1:t}$ into queries $\boldsymbol{Q}^{\boldsymbol{N}}$ and keys $\boldsymbol{K}^{\boldsymbol{N}}$ of dimension d_k, and values $\boldsymbol{V}^{\boldsymbol{N}}$. For autoregressive lip movement synthesis, multi-head self-attention is employed to capture sequence dependencies and compute attention scores for all positions using the scaled dot-product attention mechanism:

$$\mathrm{Att}(\boldsymbol{Q}^{\boldsymbol{N}}, \boldsymbol{K}^{\boldsymbol{N}}, \boldsymbol{V}^{\boldsymbol{N}}, \boldsymbol{B}^{\boldsymbol{N}}) = \mathcal{S}\left(\frac{\boldsymbol{Q}^{\boldsymbol{N}} (\boldsymbol{K}^{\boldsymbol{N}})^T}{\sqrt{d_k}} + \boldsymbol{B}^{\boldsymbol{N}}\right) \boldsymbol{V}^{\boldsymbol{N}}, \quad (7)$$

where $\mathcal{S}$ denotes the softmax function, $\boldsymbol{B}^{\boldsymbol{N}}$ is a temporal bias item. The multi-head self-attention mechanism generates a weighted context representation for each motion frame, dynamically adjusting predictions based on preceding frames. This idea enables the module to adapt to varying input lengths, ensuring consistent motion synthesis.

Content-Motion Fusion. In this module, content-oriented multi-head attention is employed to combine the outputs of the content encoder $\boldsymbol{c}_t$ and weighted motion context repre-

sentations $\boldsymbol{N}^{''}_{1:t}$ to capture various aspects of the relationship between them. To enable the attention mechanism to adjust its focus adaptively, we incorporate a relation bias to the attention computation, and the relation bias is formulated as:

$$\boldsymbol{B}^{\boldsymbol{M}}(i,j) = \begin{cases} 0 & ki \leq j < k(i+1) \\ -\infty & \text{otherwise} \end{cases}. \tag{8}$$

Then, both $\boldsymbol{c}_t$ and $\boldsymbol{N}^{''}_{1:t}$ are fed into the attention computation and $\boldsymbol{c}_t$ is transformed into two separate matrices: keys $\boldsymbol{K}^C$ and values $\boldsymbol{V}^C$, and $\boldsymbol{N}^{''}_{1:t}$ is transformed into queries $\boldsymbol{Q}^M$. Finally, we aggregate the attention-weighted fusion embedding to predict the current lip movement by decoding it back to the 3D vertex space in V dimensions through a fully connected and linear layer.

Loss Function. Once the complete 3D facial sequence is produced, the model is trained by minimizing the weighted Mean Squared Error (MSE) between the decoder outputs $\hat{\boldsymbol{F}}_{1:L}$ and the zero-posed ground truth $\boldsymbol{F}^*_{1:L}$:

$$\mathcal{L}_{\text{MSE}} = \sum_{t=1}^{L} \sum_{v=1}^{V} w_v \left\| \hat{\boldsymbol{f}}_{t,v} - \boldsymbol{f}^*_{t,v} \right\|^2, \tag{9}$$

where w_v denotes the adaptive weight introduced from Motion3DGAN (Otberdout et al. 2022). The weight offers a rough estimation of the contribution of each vertex in the lip movement synthesis based on the inverse of its Euclidean distance from the nearest vertex.

F0-Conditioned Latent Diffusion

Motion-VAE. To synthesize high-quality head poses without massive computational resources, we adopt a motion variational autoencoder, Motion-VAE, to learn a latent representation of the synthesized facial sequences and perform a conditional diffusion on the motion latent space. Inspired by MLD (Chen et al. 2023b), Motion-VAE is composed of a transformer-based architecture that consists of an encoder $\mathcal{E}$ and a decoder $\mathcal{D}$. The motion encoder $\mathcal{E}$ and decoder $\mathcal{D}$ all consist of 6 layers and 4 heads with skip connection. The encoder $\mathcal{E}$ is trained to take as input the zero-posed facial sequence $\hat{\boldsymbol{F}}_{1:L}$ and encodes the sequence into a low dimensional latent vector $\boldsymbol{z} \in R^{d_z}$. Then, the decoder $\mathcal{D}$ takes the L number of zero motion tokens as queries and a latent $\boldsymbol{z}$ as memory to reconstruct the facial sequence. To constrain the latent space as in a standard variational autoencoder, Motion-VAE is trained on facial motion reconstruction using a combination of MSE loss and Kullback-Leibler divergence between $q\left(\boldsymbol{z}|\hat{\boldsymbol{F}}_{1:L}\right) = \mathcal{N}\left(\boldsymbol{z}; \mathcal{E}_{\mu}, \mathcal{E}_{\sigma^2}\right)$ and a standard Gaussian distribution $\mathcal{N}(\boldsymbol{z}; 0, 1)$.

F0-Conditioned Latent Diffusion. Upon the latent representations of facial sequences, a conditional latent diffusion model is employed to synthesize head poses based on the F0 embedding. Unlike the previous convolutional neural networks to learn the denoising process, we adopt a transformer-based architecture with self-attention to capture long-range dependencies in sequence modeling. As depicted in Figure 2(d), the diffusion process on the motion latent space is formulated as a Markov chain as:

$$q\left(\boldsymbol{z}_t|\boldsymbol{z}_{t-1}\right) = \mathcal{N}\left(\sqrt{\alpha_t}\boldsymbol{z}_{t-1}, \left(1-\alpha_t\right) I\right), \tag{10}$$

where the constant $\alpha_t \in (0,1)$ is a hyper-parameter for sampling. We train the latent diffusion model in a classifier-free manner, which enables trading off diversity to fidelity and denosing conditionally (Tevet et al. 2023). Since the head pose variations are sensitive to F0, F0 embedding y is injected into the denosing process via adaptive instance normalization (Huang and Belongie 2017). The denoising process is implemented with conditional denoiser $\epsilon_\zeta\left(\boldsymbol{z}_t, t, \boldsymbol{y}\right)$, where ζ is the model parameters. After conditional denoising, the decoder $\mathcal{D}$ reconstructs the motion sequence from the predicted latent codes. The conditional latent diffusion is optimized by minimizing the weighted MSE loss between the reconstructed sequences $\tilde{\boldsymbol{F}}_{1:L}$ and ground truth $\boldsymbol{F}_{1:L}$.

Experiments

Dataset Construction

3DTH. Due to the scarcity of 4D face datasets with head poses, we utilize 2D talking head datasets (e.g., MEAD and HTDF) to construct our 3DTH dataset. We first resample the video as 25fps and the audio as 16kHz. We then employ EMOCA-v2 (Daněček, Black, and Bolkart 2022) to obtain the detailed FLAME parameters and compute the 3D vertices for each frame. Since the reconstruction results show occasional jitters, we introduce a sliding window mechanism to mitigate jitters. Meanwhile, we leverage the predicted facial landmarks to optimize the camera poses.

3DTH*. Except for 3DTH, we construct another 4D zero-posed face dataset, named 3DTH*, to support the training of the content-oriented lip movement synthesis. Each reconstructed motion sequence in 3DTH is transformed into zero-posed via linear blend skinning (Loper et al. 2015).

Baselines

We conduct a lip-sync comparison between our method and FaceFormer, CodeTalker, and SelfTalk. Recognizing the scarcity of visual speech animation methods addressing head poses, we compare our method with DiffPoseTalk (Sun et al. 2024) and one 2D talking face method SadTalker (Zhang et al. 2023), which incorporates head movements and utilizes a 3DMM as an intermediate facial representation.

Qualitative Evaluation

To check the lip-sync performance, we illustrate four typical frames of synthesized lip movements for specific syllables in the left partition of Figure 3. For a fair comparison, we assign the same talking style to FaceFormer and CodeTalker as conditional input, which is randomly sampled. Compared to competitors, our method synthesizes lip movements that are more accurately articulated with the speech signals and more consistent with the reference. For example, our method synthesizes better lip-sync with proper mouth closures when pronouncing bilabial consonant /b/ (i.e., "strawberry" in the upper-left case of Figure 3). It also generates accurate lip shapes for challenging long vowels that need to be pout. In contrast, others suffer from the over-smoothing problem and

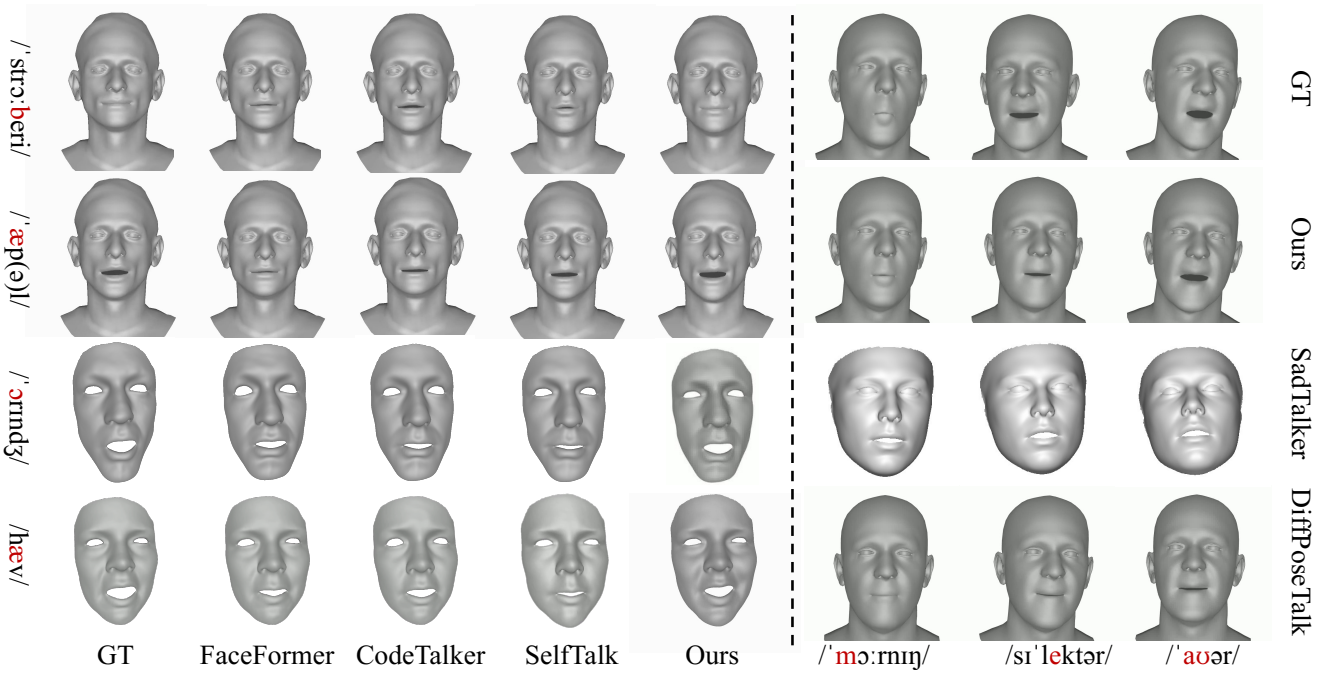

Figure 3: Qualitative comparisons about sampled lip movements and head pose from different methods on VOCASET-Test (left-upper), BIWI-Test-A (left-lower), and 3DTH-Test (right).

fail to lip-sync correctly (i.e., "orange" in the left case of Figure 3). Our method shows minimal errors, especially in capturing the lip movements for the short vowel /æ/, which requires a wide mouth opening. This superiority is primarily due to effectively incorporating the language content, enhancing robustness to cross-modal uncertainty.

Additionally, the right partition of Figure 3 illustrates three representative frames of the synthesized animation, which simultaneously captures both lip movements and head poses. In comparison to other methods, our method generates plausible head poses based on the input speech while maintaining synchronized lip movements. It demonstrates that incorporating F0 as a conditioning parameter enhances the naturalness and coherence of the synthesized head poses.

Quantitative Evaluation

To quantitatively evaluate lip synchronization, we employ two established metrics (e.g., lip vertex error (LVE) and upper face dynamics deviation (FDD)) and report the average across all sequences. Since the benchmark datasets (e.g., BIWI-Test and VOCASET-Test) are all in a neutral head pose, we exclude the head pose synthesis for comparison. As shown in Table 1, our method achieves lower errors than current state-of-the-art approaches, demonstrating its ability to generate more accurate lip-synchronized movements by leveraging the language content. Furthermore, Table 1 highlights that our method outperforms others in terms of FDD, indicating that the language content enhances the dynamics of upper face movements as well.

As shown in Table 2, we evaluate the head pose synchronization by beat alignment (BA) (Sun et al. 2024) and compute a diversity score following Diffusion motion (Ren et al. 2023). The BA calculation is revised to evaluate the synchronization of the detected head poses between the synthesized results and the corresponding ground truth. Our method surpasses all others across all metrics, achieving superior head pose beat alignment and diversity.

Methods	BIWI		VOCASET	
	LVE ↓ ($10^{-4}mm$)	FDD ↓ ($10^{-5}mm$)	LVE ↓ ($10^{-5}mm$)	FDD ↓ ($10^{-5}mm$)
FaceFormer	5.3077	4.6408	4.1090	1.5250
CodeTalker	4.7914	4.1170	3.9445	1.4812
SelfTalk	4.2485	3.5761	3.2238	0.9941
Ours	**3.8430**	**3.1934**	**2.8246**	**0.8649**

Table 1: Quantitative comparison about lip movement on BIWI-Test-A and VOCASET-Test. ↓ denotes the lower the better and vice versa. The best results are marked in bold.

Ablation Study

Impact of the Perturbation in Content Encoding. To evaluate the effectiveness of the information perturbation, we first directly remove all signal processing functions and then remove them successively to evaluate the lip-sync performance. As shown in Table 3, all processing functions offer significant benefits in enhancing the lip-sync performance by ensuring the processed speech retains its language content. Compared with other functions, shifting formants proves to be particularly effective in altering the perceived

Methods	Beat Alignment ↑	Diversity ↑
SadTalker	0.236	0.788
DiffPoseTalk	0.295	1.169
Head pose codebook	0.291	0.648
Ours without F0	0.240	0.427
Ours without F0 normalization	0.218	1.144
Ours	**0.306**	**1.182**

Table 2: Quantitative comparison about head pose and ablation study about F0 embedding on 3DTH-Test.

Methods	LVE	FDD
Feature from Wav2vec2	4.5663	1.4533
Feature from HuBERT	4.2985	1.1216
Ours without perturbation	5.9245	1.6257
Ours without pitch randomization	3.4445	0.9525
Ours without formant shifting	7.5441	1.9875
Ours without random frequency shaping	3.1090	1.0117
Ours	**2.8246**	**0.8649**

Table 3: Ablation study about the content encoder and perturbation functions on VOCASET-Test.

identity and voice characteristics of the audio while maintaining the integrity of the language content.

Impact of the Content Encoder in Lip-sync. Given the widespread use of Wav2vec2 (Baevski et al. 2020) and HuBERT (Hsu et al. 2021) for predicting lip movements, we replace the content encoder with these encoding strategies, applying them successively to perturbed speech signals. As shown in Table 3, our method achieves lower scores in both LVE and FDD, indicating the effectiveness of the content encoder in lip-sync. Additionally, an alternative approach involves using automatic speech recognition to extract the text embedding from perturbed speech signals as a substitute for lip movement synthesis. However, the text embedding ignores the temporal information in the waveform, which is necessary to achieve synchronized lip movement prediction.

Impact of F0 Embedding in Head Pose Synthesis. Inspired by AdaMesh (Chen et al. 2023a), we employ a VQ-VAE (Van Den Oord, Vinyals et al. 2017) model to cast head pose synthesis as a code query task within a learned codebook. Specifically, we combine the encoded lip movements and F0 embedding via a transformer and then input the result into the learned codebook for head pose prediction. Additionally, we omit F0 and feed a re-encoded content embedding into the latent diffusion model. As presented in Table 2, the head pose codebook achieves a BA comparable to our method, but our approach demonstrates superior diversity. Although removing F0 normalization increases head pose diversity, it also compromises the model's ability to generalize F0 variations across different speakers, leading to less consistent performance in head pose synthesis.

Methods	VOCASET-Test		3DTH-Test	
	Lip-sync	Realism	Lip-sync	Realism
Ours vs GT	43.21	47.24	46.17	49.71
Ours vs FaceFormer	68.32	62.88	-	-
Ours vs CodeTalker	51.12	70.02	-	-
Ours vs SelfTalk	49.24	63.17	-	-
Ours vs SadTalker	-	-	62.40	66.78
Ours vs DiffPoseTalk	-	-	55.22	51.27

Table 4: User study about lip-sync and animation realism. The score represents our preference over others in %.

User Study

We adopt A/B tests for each comparison, i.e., ours vs. competitor, and take the random order. In each A/B test, the participant watches the animation videos, listens to the audio clips, and answers the questions. For the VOCASET-Test and 3DTH-Test datasets, we select 20 samples from each dataset for each of the 7 comparison types. As a result, 80 A vs. B pairs (20 samples and 4 comparisons) are collected for VOCASET-Test. Meanwhile, we apply the same settings and collect another 60 A vs. B pairs for 3DTH-Test. We invite 20 participants with normal vision and hearing abilities for the user study and ensure each participant engages in all 7 types of comparisons. The group without head poses includes FaceFormer, CodeTalker, SelfTalker, our method in **zero-posed**, and the ground truth. The group with head poses involves SadTalker, DiffPoseTalk, our method, and the ground truth. The quantitative evaluation is listed in Table 4. Interestingly, our method even approximates the ground truth in animation realism, indicating that it models the data distribution based on the decomposed prosody elements, compensating for the reconstruction error. Our method has significantly outperformed others for head poses and achieves a decent motion reconstruction ability compared with the ground truth.

Conclusion

This paper emphasizes the necessity of considering speech prosody in 3D visual speech animation. We conduct extensive observations demonstrating that language content and F0 are correlated with lip movements and head poses. Based on these findings, we propose a content-oriented fusion decoder to synthesize lip movements by emphasizing content-related features. For head pose synthesis, we propose an F0-conditioned latent diffusion model to learn a probabilistic mapping from the F0 to head poses. Extensive experiments show that our method outperforms state-of-the-art methods in terms of lip-sync accuracy and head pose plausibility.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grants 62272134 and 62072141 and in part by the National Natural Science Foundation of Shandong Province under Grant ZR2024QF136.

References

Aksan, E.; Kaufmann, M.; and Hilliges, O. 2019. Structured Prediction Helps 3D Human Motion Modelling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 7144–7153.

Baevski, A.; Zhou, Y.; Mohamed, A.; and Auli, M. 2020. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. In *Proceedings of the Conference on Neural Information Processing Systems*, 12449–12460.

Chen, L.; Bao, W.; Lei, S.; Tang, B.; Wu, Z.; Kang, S.; and Huang, H. 2023a. AdaMesh: Personalized Facial Expressions and Head Poses for Speech-Driven 3D Facial Animation. *arXiv preprint arXiv:2310.07236*.

Chen, X.; Jiang, B.; Liu, W.; Huang, Z.; Fu, B.; Chen, T.; and Yu, G. 2023b. Executing your Commands via Motion Diffusion in Latent Space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18000–18010.

Choi, H.-S.; Lee, J.; Kim, W.; Lee, J.; Heo, H.; and Lee, K. 2021. Neural Analysis and Synthesis: Reconstructing Speech from Self-Supervised Representations. In *Proceedings of the Conference on Neural Information Processing Systems*, 16251–16265.

Cudeiro, D.; Bolkart, T.; Laidlaw, C.; Ranjan, A.; and Black, M. J. 2019. Capture, Learning, and Synthesis of 3D Speaking Styles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10101–10111.

Daněček, R.; Black, M. J.; and Bolkart, T. 2022. EMOCA: Emotion Driven Monocular Face Capture and Animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20311–20322.

Fan, Y.; Lin, Z.; Saito, J.; Wang, W.; and Komura, T. 2022. FaceFormer: Speech-Driven 3D Facial Animation with Transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18770–18780.

Fanelli, G.; Gall, J.; Romsdorfer, H.; Weise, T.; and Van Gool, L. 2010. A 3-D Audio-Visual Corpus of Affective Communication. *IEEE Transactions on Multimedia*, 12(6): 591–598.

Graf, H. P.; Cosatto, E.; Strom, V.; and Huang, F. J. 2002. Visual prosody: facial movements accompanying speech. In *Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition*, 396–401.

Henter, G. E.; Alexanderson, S.; and Beskow, J. 2020. MoGlow: Probabilistic and controllable motion synthesis using normalising flows. *ACM Transactions on Graphics*, 39(6): 1–14.

Hsu, W.-N.; Bolte, B.; Tsai, Y.-H. H.; Lakhotia, K.; Salakhutdinov, R.; and Mohamed, A. 2021. HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29: 3451–3460.

Huang, X.; and Belongie, S. 2017. Arbitrary Style Transfer in Real-time with Adaptive Instance Normalization. In *Proceedings of the IEEE International Conference on Computer Vision*, 1501–1510.

Loper, M.; Mahmood, N.; Romero, J.; Pons-Moll, G.; and Black, M. J. 2015. SMPL: A Skinned Multi-Person Linear Model. *ACM Transactions on Graphics*, 34(6): 1–16.

Mao, W.; Liu, M.; Salzmann, M.; and Li, H. 2019. Learning Trajectory Dependencies for Human Motion Prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9489–9497.

Ning, Z.; Xie, Q.; Zhu, P.; Wang, Z.; Xue, L.; Yao, J.; Xie, L.; and Bi, M. 2023. Expressive-VC: Highly Expressive Voice Conversion with Attention Fusion of Bottleneck and Perturbation Features. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 1–5.

Otberdout, N.; Ferrari, C.; Daoudi, M.; Berretti, S.; and Del Bimbo, A. 2022. Sparse to Dense Dynamic 3D Facial Expression Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20385–20394.

Peng, Z.; Luo, Y.; Shi, Y.; Xu, H.; Zhu, X.; Liu, H.; He, J.; and Fan, Z. 2023a. SelfTalk: A Self-Supervised Commutative Training Diagram to Comprehend 3D Talking Faces. In *Proceedings of the ACM International Conference on Multimedia*, 5292–5301.

Peng, Z.; Wu, H.; Song, Z.; Xu, H.; Zhu, X.; He, J.; Liu, H.; and Fan, Z. 2023b. EmoTalk: Speech-Driven Emotional Disentanglement for 3D Face Animation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 20687–20697.

Press, O.; Smith, N.; and Lewis, M. 2021. Train Short, Test Long: Attention with Linear Biases Enables Input Length Extrapolation. In *International Conference on Learning Representations*.

Ren, Z.; Pan, Z.; Zhou, X.; and Kang, L. 2023. Diffusion Motion: Generate Text-Guided 3D Human Motion by Diffusion Model. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 1–5. IEEE.

Richard, A.; Zollhöfer, M.; Wen, Y.; De la Torre, F.; and Sheikh, Y. 2021. MeshTalk: 3D Face Animation from Speech using Cross-Modality Disentanglement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1173–1182.

Sun, Z.; Lv, T.; Ye, S.; Lin, M.; Sheng, J.; Wen, Y.-H.; Yu, M.; and Liu, Y.-j. 2024. Diffposetalk: Speech-driven stylistic 3d facial animation and head pose generation via diffusion models. *ACM Transactions on Graphics*, 43(4): 1–9.

Tevet, G.; Raab, S.; Gordon, B.; Shafir, Y.; Cohen-or, D.; and Bermano, A. H. 2023. Human Motion Diffusion Model. In *International Conference on Learning Representations*.

Tian, Y.; Shi, J.; Li, B.; Duan, Z.; and Xu, C. 2018. Audio-Visual Event Localization in Unconstrained Videos. In *Proceedings of the European Conference on Computer Vision*, 247–263.

Van Den Oord, A.; Vinyals, O.; et al. 2017. Neural discrete representation learning. In *Proceedings of the Conference on Neural Information Processing Systems*, 6309–6318.

Wang, K.; Wu, Q.; Song, L.; Yang, Z.; Wu, W.; Qian, C.; He, R.; Qiao, Y.; and Loy, C. C. 2020. MEAD: A Large-scale Audio-visual Dataset for Emotional Talking-face Generation. In *Proceedings of the European Conference on Computer Vision*, 700–717.

Xing, J.; Xia, M.; Zhang, Y.; Cun, X.; Wang, J.; and Wong, T.-T. 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.

Ye, Z.; He, J.; Jiang, Z.; Huang, R.; Huang, J.; Liu, J.; Ren, Y.; Yin, X.; Ma, Z.; and Zhao, Z. 2023. GeneFace++: Generalized and Stable Real-Time Audio-Driven 3D Talking Face Generation. *arXiv preprint arXiv:2305.00787*.

Yehia, H. C.; Kuratate, T.; and Vatikiotis-Bateson, E. 2002. Linking facial animation, head motion and speech acoustics. *Journal of phonetics*, 30(3): 555–568.

Zhang, C.; Zhao, Y.; Huang, Y.; Zeng, M.; Ni, S.; Budagavi, M.; and Guo, X. 2021a. FACIAL: Synthesizing Dynamic Talking Face with Implicit Attribute Learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3867–3876.

Zhang, M.; Cai, Z.; Pan, L.; Hong, F.; Guo, X.; Yang, L.; and Liu, Z. 2022. MotionDiffuse: Text-Driven Human Motion Generation with Diffusion Model. *arXiv preprint arXiv:2208.15001*.

Zhang, W.; Cun, X.; Wang, X.; Zhang, Y.; Shen, X.; Guo, Y.; Shan, Y.; and Wang, F. 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.

Zhang, Z.; Li, L.; Ding, Y.; and Fan, C. 2021b. Flow-guided One-shot Talking Face Generation with a High-resolution Audio-visual Dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3661–3670.