GLDiTalker: Speech-Driven 3D Facial Animation with Graph Latent Diffusion Transformer

Yihong Lin[1*], **Zhaoxin Fan**[2,3*], **Xianjia Wu**[4], **Lingyu Xiong**[1], **Xiandong Li**[4,5†], **Wenxiong Kang**[1†], **Liang Peng**[4], **Songju Lei**[5] and **Huang Xu**[4]
[1]South China University of Technology
[2] Beijing Advanced Innovation Center for Future Blockchain and Privacy Computing, School of Artificial Intelligence, Beihang University
[3]Hangzhou International Innovation Institute, Beihang University
[4]Huawei Cloud
[5]Nanjing University
lxdphys@smail.nju.edu.cn, auwxkang@scut.edu.cn

Abstract

Speech-driven talking head generation is a critical yet challenging task with applications in augmented reality and virtual human modeling. While recent approaches using autoregressive and diffusion-based models have achieved notable progress, they often suffer from modality inconsistencies, particularly misalignment between audio and mesh, leading to reduced motion diversity and lip-sync accuracy. To address this, we propose GLDiTalker, a novel speech-driven 3D facial animation model based on a Graph Latent Diffusion Transformer. GLDiTalker resolves modality misalignment by diffusing signals within a quantized spatiotemporal latent space. It employs a two-stage training pipeline: the Graph-Enhanced Quantized Space Learning Stage ensures lip-sync accuracy, while the Space-Time Powered Latent Diffusion Stage enhances motion diversity. Together, these stages enable GLDiTalker to generate realistic, temporally stable 3D facial animations. Extensive evaluations on standard benchmarks demonstrate that GLDiTalker outperforms existing methods, achieving superior results in both lip-sync accuracy and motion diversity.

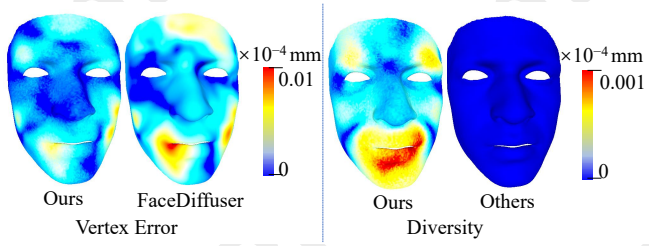

Figure 1: The left figures show the lip vertex error of GLDiTalker and FaceDiffuser and the right figures show the diversity of GLDiTalker and other methods, including FaceFormer, CodeTalker, TalkingStyle and FaceDiffuser. "Vertex Error" refers to the average L_2 loss between prediction and groundtruth. "Diversity" refers to the statistical mean of the motion differences between corresponding frames in multiple generation results within the same sequence. It is obvious that GLDiTalker has a significant advantage in both generation quality and motion diversity.

1 Introduction

Speech-driven talking head generation aims to synthesize realistic and synchronized facial motion from speech, with applications in assistive technologies, such as virtual reality and augmented reality [Ping *et al.*, 2013; Wohlgenannt *et al.*, 2020]. Despite its potential, the task is challenging due to the many-to-many mapping between audio and facial motion, requiring a balance between accurate lip synchronization and natural expression generation.

To achieve realistic speech-driven talking head generation, various methods have been proposed. Traditional approaches [Edwards *et al.*, 2016; Taylor *et al.*, 2012; Xu *et al.*, 2013] rely on manual or rule-based designs, which are often costly and inflexible. In contrast, deep learning methods [Karras *et al.*, 2017; Peng *et al.*, 2023a] demonstrate stronger adaptability and have made significant progress. VOCA [Cudeiro *et al.*, 2019] employs a CNN-based framework to generate cross-identity facial animations, while MeshTalk [Richard *et al.*, 2021] leverages U-Net to enhance expression diversity by modeling latent facial expression spaces. Transformer-based models such as FaceFormer [Fan *et al.*, 2022] improve temporal and cross-modal alignment through mechanisms like Biased Cross-Modal Multi-Head Attention. Similarly, CodeTalker [Xing *et al.*, 2023] introduces a VQ-VAE-based motion prior to reduce the uncertainty in audio-to-motion mapping. Despite these advancements, existing methods share common limitations. Most approaches rely on deterministic models, which generate identical 3D facial motions for the same input conditions. This contradicts real-world behavior, where non-verbal facial cues are inherently nondeterministic [Ng *et al.*, 2022]. Even when the same individual repeats the same sentence, subtle variations in facial motion naturally occur, reflecting the complexity of human

communication. By failing to capture this variability, current methods often suffer from modality inconsistencies, particularly misalignment between audio and mesh. These issues lead to reduced motion diversity and compromised lip-sync accuracy, ultimately undermining the realism and expressiveness of generated animations.

To tackle the issues, we propose GLDiTalker, a Graph Latent Diffusion Transformer for audio-driven talking head generation. GLDiTalker first encodes facial motion into a spatio-temporal quantized latent space, effectively addressing the audio-mesh modality misalignment to ensure accurate lip synchronization. It then diffuses the signal within this latent quantized spatio-temporal space, introducing a controlled degree of stochasticity to enhance motion diversity. Existing methods excel in either lip-sync accuracy or motion diversity but struggle with both due to audio-mesh modality misalignment. In contrast, GLDiTalker resolves this issue and achieves superior performance in both aspects, as shown in Fig. 1.

To achieve this goal, GLDiTalker adopts a carefully designed quantized space-time diffusion training pipeline, which consists of two complementary stages: the Graph Enhanced Quantized Space Learning Stage and the Space-Time Powered Latent Diffusion Stage. In the first stage, GLDiTalker focuses on encoding facial motion into a spatio-temporal quantized latent space to ensure accurate lip synchronization. Unlike existing approaches that primarily emphasize temporal facial motion, our method simultaneously captures temporal correlations and spatial connectivities between vertices. To this end, we propose a novel Spatial Pyramidal SpiralConv Encoder, which extracts multi-scale features using spiral convolution across multiple levels. This design enables GLDiTalker to effectively integrate diverse levels of semantic information, laying a solid foundation for accurate and realistic lip synchronization. In the second stage, GLDiTalker employs a diffusion model to iteratively add and remove noise to the latent quantized facial motion features obtained from the first stage. This process introduces a controlled degree of stochasticity, enhancing motion diversity while maintaining the integrity of the learned features. Together, these two stages ensure that GLDiTalker achieves a balance between precise lip-sync and rich motion diversity, overcoming the limitations of previous methods.

Extensive experiments on existing datasets demonstrate that our method significantly outperforms previous state-of-the-art approaches in both lip-sync accuracy and motion diversity, effectively addressing the limitations of current techniques. Our contribution can be summarized as:

- We introduce GLDiTalker, a novel framework that addresses the conflict between achieving accurate lip synchronization and diverse facial motion caused by audio-mesh modality misalignment.
- GLDiTalker features a two-stage design: a Graph Enhanced Quantized Space Learning Stage for encoding spatio-temporal facial motion and a Space-Time Powered Latent Diffusion Stage for enhancing motion diversity.
- Experiments on VOCASET and BIWI datasets demonstrate that GLDiTalker outperforms state-of-the-art methods in both lip-sync accuracy and motion diversity.

2 Related Work

2.1 Speech-Driven 3D Facial Animation

Speech-driven talking face generation has attracted significant attention and achieved remarkable progress in recent years [Karras *et al.*, 2017; Richard *et al.*, 2021; Fan *et al.*, 2022; Xing *et al.*, 2023; Peng *et al.*, 2023a; Wu *et al.*, 2023; Stan *et al.*, 2023; Shen *et al.*, 2025b]. Early methods, such as VOCA [Cudeiro *et al.*, 2019], introduce CNN-based frameworks for cross-identity face animation, while later works like Mimic [Fu *et al.*, 2024] and EmoTalk [Peng *et al.*, 2023b] focus on decoupling speech content, style, and emotion. Transformer-based approaches such as FaceFormer [Fan *et al.*, 2022] improve audio-visual mapping through attention mechanisms, and CodeTalker [Xing *et al.*, 2023] reduces cross-modal uncertainty using discrete motion priors. Recently, FaceDiffuser [Stan *et al.*, 2023] applies diffusion mechanisms to enhance the diversity of facial motion generation. Although these methods address various challenges, they still struggle to simultaneously achieve high accuracy and diversity due to limitations in modality alignment and inherent model constraints. To overcome these issues, we propose GLDiTalker, which introduces a quantized space-time diffusion training pipeline to better balance precision and diversity in speech-driven talking face generation.

2.2 Diffusion in Talking Head Animation

Inspired by the diffusion processes in non-equilibrium thermodynamics, [Sohl-Dickstein *et al.*, 2015] proposes diffusion, which differs from generative models like GAN [Goodfellow *et al.*, 2020], VAE [Kingma and Welling, 2014], and Flow [Kingma and Dhariwal, 2018]. Diffusion is a Markov process capable of generating high-quality images from Gaussian noise, exhibiting a certain degree of stochasticity [Shen and Tang, 2024; Shen *et al.*, 2025a]. Recently, many 2D talking head generation methods based on diffusion models have achieved highly realistic results [Park *et al.*, 2022; Shen *et al.*, 2022; Zhang *et al.*, 2023; Shen *et al.*, 2023; Xu *et al.*, 2024]. For example, using audio as a condition, DREAM-Talk [Zhang *et al.*, 2023] designs a two-stage diffusion-based framework that generates diverse expressions while maintaining accurate audio-lip synchronization. DiffTalk [Shen *et al.*, 2023] further incorporates reference images and facial landmarks, enabling personality-aware synthesis and generalization across identities without fine-tuning. Similarly, VASA-1 [Xu *et al.*, 2024] employs diffusion in a latent space to model overall facial dynamics and head movements, effectively capturing subtle facial expressions and head motions. In the context of 3D facial animation, FaceDiffuser [Stan *et al.*, 2023] is the first to explicitly apply diffusion for speech-driven 3D facial motion generation. While diffusion methods inherently enhance diversity due to their stochastic nature, they often struggle to achieve high accuracy, particularly in terms of lip synchronization and motion precision. This trade-off between diversity and accuracy remains a significant challenge. To address this, we

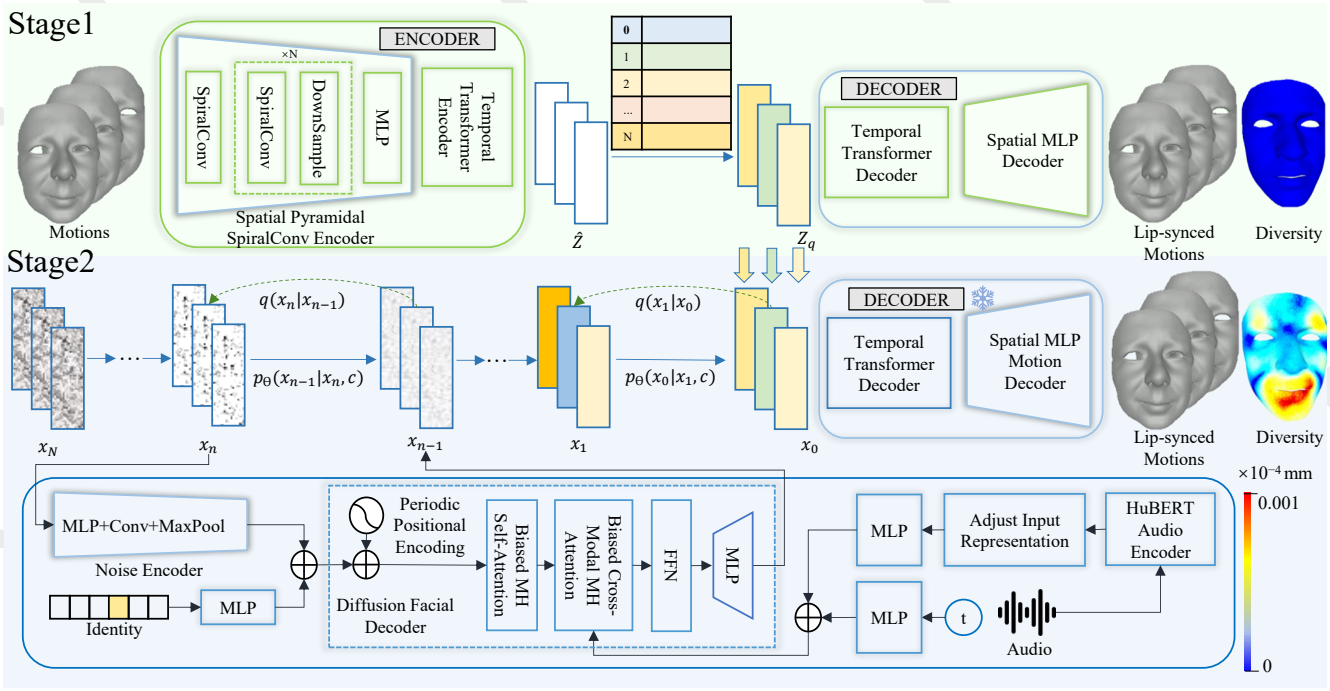

Figure 2: Overview of the proposed GLDiTalker. In order to enhance the lip-sync accuracy, Stage 1 employs the Spatial Pyramidal SpiralConv Encoder and temporal transformer encoder to sequentially process the input motions to obtain discrete latent facial motions. To enhance the motion diversity, Stage 2 then utilizes a diffusion-based iterative denoising technique to synthesize the latent facial motions from input speech and speaker identity, which is later decoded to a motion mesh sequence by the pretrained frozen decoder of the first stage. "⊕" denotes addition and "N" denotes N identical blocks, each consisting of a SpiralConv layer and a Downsample layer. The heatmaps show the statistical mean of the motion differences between corresponding frames in multiple generation results within the same sequence.

propose GLDiTalker, which leverages a quantized space-time diffusion training pipeline to balance diversity and accuracy.

3 Method

3.1 Overview

Problem Definition

The facial motion $M_{1:T} = (m_1, m_2, ..., m_T)$ is a sequence of vertex displacements $m_i \in \mathbb{R}^{V\times 3}$ applied to a neutral template face mesh $f \in \mathbb{R}^{V\times 3}$, which consists of V vertices. The audio $A_{1:T} = (a_1, a_2, ..., a_T)$ is a sequence of speech snippets, where each $a_i \in \mathbb{R}^C$ contains C samples aligned with the corresponding motion m_i. Our objective is to predict a mesh sequence $\hat{M}_{1:T}$ based on the audio $A_{1:T}$ and the talking style represented by one-hot vectors $S = (s_1, s_2, ..., s_I)$, where I denotes the number of speaking styles. The final face animations are obtained by summing the neutral template and the predicted motion, expressed as $\hat{F}_{1:T} = \hat{m}_{1+f}, ..., \hat{m}_{T+f}$. The output sequence $\hat{F}_{1:T}$ is expected to achieve accurate lip synchronization with the input audio while exhibiting diverse and expressive facial motions that reflect different speaking styles.

Quantized Space-Time Diffusion Training Pipeline

To solve the problem of failing to simultaneously achieve high lip-sync accuracy and motion diversity caused by modality misalignment, GLDiTalker follows a two-stage pipeline that consists of a Graph Enhanced Quantized Space Learning Stage and a Space-Time Powered Latent Diffusion Stage, as illustrated in Fig. 2. In the first stage, GLDiTalker focuses on quantization to enhance lip-sync accuracy. Specifically, it employs a spatio-temporal VQ-VAE [Van Den Oord *et al.*, 2017] based on graph convolution and transformer to model facial motion into a quantized facial motion prior, ensuring precise alignment between audio and facial motion. In the second stage, the focus shifts to diffusion for generating diverse motions. A transformer-based denoising network is adopted for the reverse diffusion process, which iteratively transforms a standard Gaussian distribution into the facial motion prior. This process is conditioned on audio, speaker identity, and the diffusion step, enabling the generation of realistic and expressive facial animations. The animation results can be directly obtained from the denoising outputs by using the decoder from the first stage.

Next, we introduce the Graph Enhanced Quantized Space Learning Stage and the Space-Time Powered Latent Diffusion Stage in detail, respectively.

3.2 Stage 1: Graph Enhanced Quantized Space Learning for Lip-sync Accuracy Improvement

Existing methods have limited results on lip-sync, mostly limited by the ability to model the explicit space or the accumulation of errors due to autoregression. To improve the robustness to cross-modal uncertainty, GLDiTalker proposes a

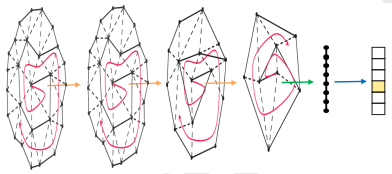

Figure 3: Architecture of Spatial Pyramidal SpiralConv Encoder. The spiral lines represent spiral convolution, the orange arrows represent downsampling, the green arrow represents flatten, and the blue arrow represents fusion by MLP.

Graph Enhanced Quantized Space Learning Stage inspired by VQ-VAE to model the facial motion into a quantized motion prior, as illustrated in the first stage in Fig. 2. A graph based encoder is trained with both the temporal information dependency and the spatial connectivity of mesh vertices taken into account. Specifically, spatial and temporal information are serially processed to get latent feature $\hat{Z} \in \mathbb{R}^{T\times H\times C}$ with a novel designed Spatial Pyramidal SpiralConv Encoder E^V following by a temporal transformer encoder E^T:

$$E(M_{1:T}) = E^T(E^V(M_{1:T})) \rightarrow \hat{Z}_{1:T}, \tag{1}$$

where H and C are the number of face components and channels, respectively.

The Spatial Pyramidal SpiralConv Encoder consists of SpiralConv layers, downsampling layers and a fusion layer, as shown in Fig. 3. SpiralConv [Gong *et al.*, 2019], a graph-based convolution operator is introduced to process mesh data, which determines the convolution centre and produces a sequence of enumerated centre vertices based on adjacency, followed by the 1-ring vertices, the 2-ring vertices, and so on, until all vertices containing k rings are included. SpiralConv determines the adjacency in the following way:

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

where $\mathcal{NH}(V)$ selects all vertices in the neighborhood of any vertex in set V. SpiralConv generates spiral sequences only once and explicitly encodes local information. It adopts a fully-connected layer for feature fusion:

$$SpiralConv(v) = W(f(k{-}disk(v))) + b, \tag{3}$$

where W and b are learnable weights and bias.

Then the SpiralConv layer and the downsampling layer are novelly composed into a Spiral Block, which reduces the number of vertices and increases the feature channels during feature extraction. After stacking multiple Spiral Blocks, the resulting graphs with a small scale are flattened and fed to the fusion layer (an MLP layer).

The Spatial Pyramidal SpiralConv Encoder is capable of extracting multi-scale features from the input graph by performing SpiralConv at multiple scales. This module enables the network to integrate semantic-level information at different levels. Lower-level pyramids can capture more detailed information, while higher-level pyramids focus more on abstract and semantic information. Pyramid network can effectively reduce the computational cost and improve the efficiency of the model in comparison to a single-scale network.

After getting the latent feature $\hat{Z} \in \mathbb{R}^{T\times C}$, the discrete facial motion latent sequence $Z_q \in \mathbb{R}^{T\times C}$ is obtained by an element-wise quantization function $Q(\cdot)$, which calculates by a nearest neighbour look-up in codebook $\mathcal{Z} = \left\{z_m \in \mathbb{R}^C\right\}_{m=1}^{M}$:

$$Z_q = Q(\hat{Z}) \rightarrow \arg\min \left\|\hat{z}_t - z_m\right\|_2, z_m \in \mathcal{Z}. \tag{4}$$

Finally, temporal transformer decoder D^T and spatial MLP decoder D^V are adopted for self-reconstruction:

$$\hat{M}_{1:T} = D(Z_q) = D^V(D^T(Z_q)). \tag{5}$$

Constraints

In this stage, we train with the following loss functions:

$$L_{stage1} = \lambda_{rec1} L_{rec1} + \lambda_{quant} L_{quant}, \tag{6}$$

where $\lambda_{rec1} = \lambda_{quant} = 1$.

Motion Reconstruction Loss calculates the L1 loss between the predicted animation sequence $\hat{M}_{1:T}$ and the reference facial motion $M_{1:T}$:

$$L_{rec1} = \left\|\hat{M}_{1:T} - M_{1:T}\right\|_1. \tag{7}$$

Quantization Loss contains two intermediate code-level losses that update the codebook items by reducing the MSE distance between the codebook $\mathcal{Z}$ and latent features $\hat{Z}$:

$$L_{quant} = \left\|sg(\hat{Z}) - Z_q\right\|_2^2 + \beta \left\|\hat{Z} - sg(Z_q)\right\|_2^2, \tag{8}$$

where $sg(\cdot)$ stands for the stop-gradient operator, which is defined as identity in the forward computation with zero partial derivatives; β denotes a weighted hyperparameter, which is 0.25 in all our experiments.

3.3 Stage 2: Space-Time Powered Latent Diffusion for Motion Diversity Enhancement

To enhance the motion diversity that neglected by the deterministic models, GLDiTalker designs a Space-Time Powered Latent Diffusion Stage to introduce the quantized facial motion prior obtained from the first stage and subsequently train a speech-conditioned diffusion-based model on top of it. During the forward process, a Markov chain $q(x_n|x_{n-1})$ for $n \in \{1, ..., N\}$ gradually introduces Gaussian noise to the clean data sample Z_q (denoted as x_0), resulting in a standard normal distribution $q(x_N|x_0)$:

$$q(x_N|x_0) = \prod_{n=1}^{N} q(x_n|x_{n-1}), \tag{9}$$

where N is the diffusion step.

On the contrary, the distribution $p(x_{n-1}|x_n)$ is employed to reconstruct the clean data sample x_0 from the noise x_N during the backward process. Conditioning on the audio $A_{1:T}$, one-hot speaker style embedding s_i and diffusion step

n, we apply a Markov chain $p(x_{n-1}|x_n, A_{1:T}, s_i, n)$ to sequentially transform the distribution $p(x_N)$ into $p(x_0)$:

$$p(x_0|A_{1:T}, s_i, N) = \prod_{n=1}^{N} p(x_{n-1}|x_n, A_{1:T}, s_i, n) \cdot p(x_N), \quad (10)$$

where $p(x_N) = \mathcal{N}(0, I)$. In practice, we use a denoising network G that directly predicts the clean sample $\hat{x}_0$:

$$\hat{x}_0 = G(x_N, A_{1:T}, s_i, N). \quad (11)$$

As the second stage shown in Fig. 2, the architecture of the denoising network G consists of five modules: (1) a noise encoder E_{noise}; (2) a style encoder E_s; (3) an audio encoder E_a; (4) a denoising step encoder E_d and (5) a diffusion facial decoder D.

Noise Encoder E_{noise} reduces the dimension of the latent features, thereby retaining the most informative features while reducing the computational complexity of model training and inference, thus improving the efficiency and speed of the model.

Style Encoder E_s encodes the one-hot speaker style embedding $s_i \in \mathbb{R}^I$ into a latent code.

Audio Encoder E_a uses the released *hubert-base-ls960* version of the HuBERT architecture pre-trained on 960 hours of 16kHz sampled speech audio. The feature extractor, feature projection layer and the initial two layers of the encoder are frozen, while the remaining parameters are set to be trainable, which enables HuBERT to better explore the motion information from audio. Since the facial motions might be captured with a frequency different from that of the speech tokens, we adjust input representation similar to [Stan *et al.*, 2023].

Denoising Step Encoder E_d encodes the denoising step n into a latent code.

Diffusion Facial Decoder D is a transformer decoder that produces the final animation latent feature sample. The decoder process can be abstracted as follows:

$$\begin{aligned}\hat{x}_0 &= D(E_{noise}(x_N), E_a(A_{1:T}), E_s(s_i), E_d(N)) \\ &= f(CA(SA(E_n(x_N) + E_s(s_i)), E_a(A_{1:T}) + E_d(N))),\end{aligned} \quad (12)$$

where $CA(q, kv)$, $SA(\cdot)$ and $f(\cdot)$ mean Biased Cross-Modal Multi-Head Attention, Biased Causal Multi-Head Self-Attention and feed forward network, respectively; q denotes query input features and kv denotes key and value input features. The alignment mask [Fan *et al.*, 2022] in CA makes the motion features only attend to the speech features at the same position.

Constraints

The loss function of this stage consists of two items:

$$L_{stage2} = \lambda_{rec2} L_{rec2} + \lambda_{vel} L_{vel}, \quad (13)$$

where $\lambda_{rec2} = \lambda_{vel} = 1$.

Latent Feature Reconstruction Loss uses a Huber loss to supervise the recovered $\hat{x}_0$:

$$L_{rec2} = \|\hat{x}_0 - x_0\|_H. \quad (14)$$

Besides, Velocity Loss also employs a Huber loss to supervise inter-frame variations for smooth animation:

$$L_{vel} = \left\|(\hat{x}_0^{2:T} - \hat{x}_0^{1:T-1}) - (x_0^{2:T} - x_0^{1:T-1}))\right\|_H. \quad (15)$$

4 Experiments

4.1 Datasets and Implementations

We conduct abundant experiments on two public 3D facial datasets, BIWI [Fanelli *et al.*, 2010] and VOCASET [Cudeiro *et al.*, 2019], both of which have 4D face scans along with audio recordings.

BIWI dataset. The BIWI dataset contains 40 sentences spoken by 14 subjects. The recordings are captured at 25 fps with 23370 vertices per mesh. We follow the data splits of the previous work [Fan *et al.*, 2022] and only use the emotional data for fair comparisons. Specifically, the training set (BIWI-Train) contains 192 sentences, the validation set (BIWI-Val) contains 24 sentences, and the testing set are divided into two subsets, in which BIWI-Test-A contains 24 sentences spoken by 6 seen subjects during training and BIWI-Test-B contains 32 sentences spoken by 8 unseen subjects during training. BIWI-Test-A is used for both quantitative and qualitative evaluation and BIWI-Test-B is used for qualitative testing.

VOCASET dataset. The VOCASET dataset consists of 480 audio-visual pairs from 12 subjects. The facial motion sequences are captured at 60 fps with about 4 seconds in length. The 3D face meshes in VOCASET are registered to the FLAME topology [Li *et al.*, 2017], with 5023 vertices per mesh. Similar to [Fan *et al.*, 2022], we adopt the same training (VOCA-Train), validation (VOCA-Val) and testing (VOCA-Test) splits for qualitative testing.

4.2 Quantitative Evaluation

We adopt the following metrics for quantitative evaluation:

- **Lip vertex error (LVE)** measures the deviation of the lip vertices of the generated sequences relative to the ground truth by calculating the maximal L2 loss for each frame and averaging over all frames.
- **Facial Dynamics Deviation (FDD)** measures the deviation of the upper face motion variation of the generated sequences relative to the ground truth. The standard deviation of the elements-wise L2 norm along the temporal axis at the upper face vertices of the generated sequence and the ground truth are initially calculated. Subsequently, their differences are solved and averaged.
- **Diversity** measures the diversity of the generated facial motions from the same audio input. We calculate this metric across all samples in the BIWI-Test-A following [Ren *et al.*, 2023].

Since all subjects in BIWI-Test-A have all been seen in the training dataset, the model can learn the motion styles of the subjects in the testing set, which exhibits a relatively fixed output under speech-driven conditions. In contrast, the subjects in BIWI-Test-B and VOCASET-Test have not been included in the training dataset. Consequently, the model is only able to predict the animation based on the styles of the

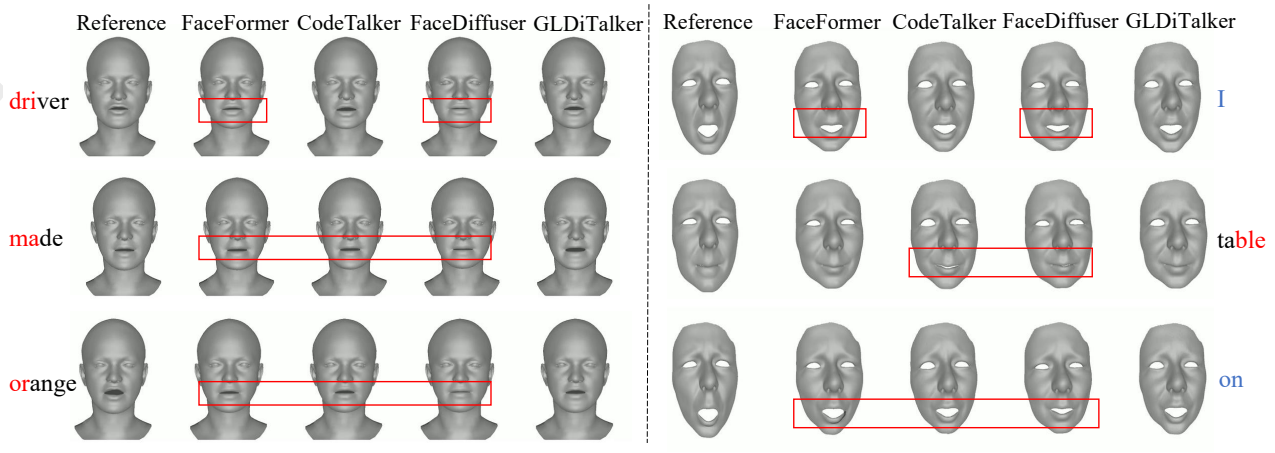

Figure 4: Qualitative Comparision on VOCASET-Test (left) and BIWI-Test-B (right).

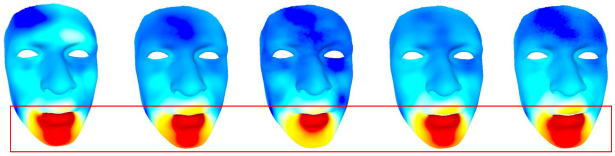

Figure 5: Standard deviation motion heatmaps of facial motion within a random sampled sequence, where dark blue means less motion variations and bright red means more motion variations.

subjects in the training set. The results may differ significantly from the ground truth, but it may be reasonable to consider without the styles. Therefore, we only quantitatively compare the performance of our GLDiTalker with state-of-the-art methods on BIWI-Test-A, as in previous work [Xing *et al.*, 2023] and [Stan *et al.*, 2023].

Quantitative evaluations are shown in Tab. 1. Similar to the superior performance of CodeTalker compared to FaceFormer, GLDiTalker outperforms FaceDiffuser in terms of lip movement and upper facial expression motion prediction (0.1383×10^{-4}mm and 0.0751×10^{-5}mm decrease in LVE and FDD, respectively) by leveraging VQ-VAE to introduce a motion prior that enhances the robustness of the cross-modal mapping from audio to 3D vertex displacements. This strongly proves the effectiveness of our quantized space-time diffusion pipeline. Although TalkingStyle stands out with the lowest LVE and FDD, our approach ranks closely behind, showcasing only a minor difference. Furthermore, the highest Diversity (8.2176×10^{-4}mm) demonstrates the superiority of our model in capturing many-to-many mappings between audio and facial motion. The latent diffusion model that starts with random noise and generates latent features through multiple iterations of denoising extremely enhances motion diversity compared to deterministic methods and explicit diffusion model method.

Methods	LVE ↓ ($\times10^{-4}$mm)	FDD ↓ ($\times10^{-5}$mm)	Diversity ↑ ($\times10^{-4}$mm))
VOCA[2019]	6.5563	8.1816	0
MeshTalk[2021]	5.9181	5.1025	0
FaceFormer[2022]	5.3077	4.6408	0
CodeTalker[2023]	4.7914	4.1170	0
FaceDiffuser[2023]	4.7823	3.9225	5.6421×10^{-5}
TalkingStyle[2024]	4.3646	3.8361	0
GLDiTalker	**4.6440**	**3.8474**	**8.2176**

Table 1: Quantitative evaluations on BIWI-Test-A.

Methods	VOCASET-Test		BIWI-Test-B	
	Realism↑	LipSync↑	Realism↑	LipSync↑
FaceFormer	3.02	3.08	3.38	3.50
CodeTalker	3.46	3.37	3.56	3.45
FaceDiffuser	2.55	2.47	3.04	2.97
GLDiTalker	**3.94**	**4.05**	**3.85**	**4.00**

Table 2: User study results.

4.3 Qualitative Evaluation

We visually compare our GLDiTalker with FaceFormer, CodeTalker and FaceDiffuser and randomly assign the same speaking style as the conditional input to make a fair comparison. Fig. 4 shows the animation results of the audio sequences from VOCASET-Test and BIWI-Test-B. We can observe that the lip movements produced by GLDiTalker are more accuracy. For vowel articulation, GLDiTalker exhibits a larger mouth opening than FaceFormer and FaceDiffuser, which is more consistent with GroundTruth. This can be observed in the mouth shape of the three examples in VOCASET-Test as well as 'I' and 'on' in BIWI-Test-B. For consonant articulation, GLDiTalker can fully close the mouth but CodeTalker and FaceDiffuser are less effective, which can be reflected by 'table' in BIWI-Test-B. Readers are recommended to watch the the Supplemental Video for more detail. In addition, we also visualize the standard deviation of facial movements of the four methods. Fig. 5 shows that our GLDiTalker outperforms others in the range of facial motion variations, illustrating the generation diversity. In addition,

Methods	LVE ↓ ($\times 10^{-4}$mm)	FDD ↓ ($\times 10^{-5}$mm)	Diversity ↑ ($\times 10^{-4}$mm)
w/o Graph Enhanced Quantized Space Learning Stage	4.7823	3.9225	5.6421×10^{-5}
w/ Spatial MLP Encoder	4.8909	4.3384	8.4251
w/ Spatial Pyramidal SpiralConv Encoder	**4.6440**	**3.8474**	**8.2176**

Table 3: Ablation study for Graph Enhanced Quantized Space Learning Stage on BIWI-Test-A.

Kernel Size	Dilation	Layer	LVE ↓ ($\times 10^{-4}$mm)	FDD ↓ ($\times 10^{-5}$mm)
5	1	4	4.9904	3.9252
9	1	2	4.9689	4.0887
9	1	4	4.7745	3.7400
9	2	3	5.1648	4.1967
9	2	4	**4.6440**	**3.8474**

Table 4: Ablation study for hyperparameters of our Spatial Pyramidal SpiralConv Encoder on BIWI-Test-A.

the right side of Fig. 1 demonstrates the highest motion diversity of GLDiTalker than the other methods. More comparisons are shown in *Supplementary Video*.

4.4 User Study

In order to evaluate our model more comprehensively, we conduct a user study to evaluate realism and lip synchronization. Specifically, a questionnaire is designed for the purpose of comparing the effectiveness of FaceFormer, CodeTalker, FaceDiffuser and GLDiTalker with ground truth on 20 randomly selected samples from BIWI-Test-B and VOCASET-Test, respectively. Twenty participants are shown the questionnaire and asked to rate on a scale of 1-5. We count the Mean Opinion Score (MOS) of all methods. Tab. 2 shows that GLDiTalker gets the highest score in both two metrics, demonstrating its superiority on semantic comprehension and natural realism over other methods.

4.5 Ablation Study

In this section, we conduct an ablation study to the impact of our proposed quantized space-time diffusion training pipeline on the quality of generated 3D talking faces. All the experiments are conducted on BIWI dataset and Tab. 3 and Tab. 4 are the results from BIWI-Test-A.

Impact of Graph Enhanced Quantized Space Learning Stage

Tab. 3 illustrates that our model achieves obvious superiority in lip-sync accuracy due to the Graph Enhanced Quantized Space Learning Stage with Spatial Pyramidal SpiralConv Encoder. The lowest LVE and FDD (4.6440×10^{-4}mm and 3.8474×10^{-5}mm in LVE and FDD, respectively) indicate the considerable impact on the overall visual quality of the generated faces and the comprehensibility of lip movements. With regard to the local connectivity of the graph, Spatial Pyramidal SpiralConv Encoder updates the representation of each mesh vertex by fusing the features with those of the neighbouring vertices. This process preserves the topology information of the vertices, thereby encouraging the model to effectively learn the representation of each vertex and the whole graph structure. In contrast, MLP is typically oriented towards regular input data with overly dense connections that do not take into account the structure of the graph.

Methods	LVE ↓ ($\times 10^{-4}$mm)	FDD ↓ ($\times 10^{-5}$mm)
w/o Noise Encoder E_{noise}	5.5239	6.6215
w Noise Encoder E_{noise}	**4.6440**	**3.8474**

Table 5: Ablation Study for Noise Encoder E_{noise} on BIWI-Test-A.

A series of ablation experiments on kernel size, dilation coefficient, and layer numbers of Spatial Pyramidal SpiralConv Encoder are provided, as shown in Tab. 4. With respect to kernel size, the data in the first and third rows indicates that a large kernel size confers a significant advantage in LVE and FDD (0.2159×10^{-4}mm and 0.1852×10^{-5}mm decrease in LVE and FDD, respectively). This may be attributed to the fact that larger convolution kernels have larger receptive field, thus enabling the integration of a greater quantity of information. Furthermore, in terms of the dilation coefficients, the third and fifth rows show that larger dilation coefficients lead to better lip synchronisation, although this is associated with a reduction in facial quality performance (0.1305×10^{-4}mm decrease in LVE and 0.1074×10^{-5}mm increase in FDD). We believe that the effect of lip synchronisation is more significant, so we retain the option of using a large dilation factor. Finally, for the number of layers, two sets of experiments are conducted to demonstrate the importance of the pyramid layer number. With a dilation factor of 1 (no dilation convolution operation), changing the number of layers from 4 to 2 leads to a decline in performance across all metrics (0.1944×10^{-4}mm and 0.3487×10^{-5}mm increase in LVE and FDD, respectively), as illustrated in the second and third rows; in the case of dilation factor of 2, modifying the number of layers from 4 to 3 also leads to worse performance (0.5208×10^{-4}mm and 0.3493×10^{-5}mm increase in LVE and FDD, respectively), as shown in the fourth and fifth rows. This suggests that an increase in the layer number can enhance feature extraction capability to comprehensively capture the intrinsic structure and regularity of the sparse vertex data effectively.

Impact of Space-Time Powered Latent Diffusion Stage

The last column of the Tab. 3 shows that using diffusion in the quantized space of Graph Enhanced Quantized Space Learning Stage leads to significant improvements in terms of motion diversity (8.2176×10^{-4}mm). With the same input, the deterministic model can only produce a fixed output, which does not reflect the many-to-many correspondence between speech and facial motions in reality. Instead, our approach uses diffusion in the quantized space, which even has an advantage of several orders of magnitude over FaceDiffuser, which adopts diffusion in the explicit space. We also conducted experiments to remove the Noise Encoder E_{noise}, which resulted in a performance degradation, as shown in Tab. 5, indicating the need of Noise Encoder.

5 Conclusion and Discussion

In conclusion, we propose GLDiTalker, a novel speech-driven 3D facial animation model that addresses the challenges of modality misalignment, lip-sync accuracy, and motion diversity in talking head generation. By introducing a two-stage pipeline, GLDiTalker ensures precise lip synchronization through the Graph-Enhanced Quantized Space Learning Stage and enhances motion diversity with the Space-Time Powered Latent Diffusion Stage. Extensive evaluations demonstrate that GLDiTalker achieves state-of-the-art performance, generating realistic, temporally stable 3D facial animations with both high lip-sync accuracy and diverse motion, making it a robust solution for speech-driven animation tasks.

Ethical Statement

Face data can be used for generating content that may jeopardize privacy. We must act responsibly by considering the aspects related to privacy and ethics.

Acknowledgements

This work was supported by the National Natural Science Foundation of China under Grant No. 62441617. It was also supported by the Beijing Natural Science Foundation under Grant No. 4254100, the Fundamental Research Funds for the Central Universities under Grant No. KG16336301, the China Postdoctoral Science Foundation under Grant No. 2024M764093, and by Beijing Advanced Innovation Center for Future Blockchain and Privacy Computing.

Contribution Statement

Xiandong Li and Wenxiong Kang are the corresponding authors. They led the planning and execution of the research and provided financial support for the project leading to this publication. Yihong Lin and Zhaoxin Fan contributed equally to this work as co-first authors. Yihong Lin and Xiandong Li proposed the ideas. In addition, Yihong Lin designed the model and the implementation of the computer code, performed the experiments and wrote the initial draft. Zhaoxin Fan analyzed the study data and made critical review, commentary and revision including pre and post publication stages. Xianjia Wu and Lingyu Xiong assisted with figure preparation. Liang Peng, Songju Lei and Huang Xu proofread the manuscript. All authors have read and approved the final manuscript.

References

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

[Edwards *et al.*, 2016] Pif Edwards, Chris Landreth, Eugene Fiume, and Karan Singh. Jali: an animator-centric viseme model for expressive lip synchronization. *ACM Transactions on graphics (TOG)*, 35(4):1–11, 2016.

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

[Fanelli *et al.*, 2010] Gabriele Fanelli, Juergen Gall, Harald Romsdorfer, Thibaut Weise, and Luc Van Gool. A 3-d audio-visual corpus of affective communication. *IEEE Transactions on Multimedia*, 12(6):591–598, 2010.

[Fu *et al.*, 2024] Hui Fu, Zeqing Wang, Ke Gong, Keze Wang, Tianshui Chen, Haojie Li, Haifeng Zeng, and Wenxiong Kang. Mimic: Speaking style disentanglement for speech-driven 3d facial animation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 1770–1777, 2024.

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

[Goodfellow *et al.*, 2020] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

[Karras *et al.*, 2017] Tero Karras, Timo Aila, Samuli Laine, Antti Herva, and Jaakko Lehtinen. Audio-driven facial animation by joint end-to-end learning of pose and emotion. *ACM Transactions on Graphics (TOG)*, 36(4):1–12, 2017.

[Kingma and Dhariwal, 2018] Durk P Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. *Advances in neural information processing systems*, 31, 2018.

[Kingma and Welling, 2014] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *International Conference on Learning Representations (ICLR)*, 2014.

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

[Ng *et al.*, 2022] Evonne Ng, Hanbyul Joo, Liwen Hu, Hao Li, Trevor Darrell, Angjoo Kanazawa, and Shiry Ginosar. Learning to listen: Modeling non-deterministic dyadic facial motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20395–20405, 2022.

[Park *et al.*, 2022] Se Jin Park, Minsu Kim, Joanna Hong, Jeongsoo Choi, and Yong Man Ro. Synctalkface: Talking face generation with precise lip-syncing via audio-lip memory. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 2062–2070, 2022.

[Peng *et al.*, 2023a] Ziqiao Peng, Yihao Luo, Yue Shi, Hao Xu, Xiangyu Zhu, Hongyan Liu, Jun He, and Zhaoxin Fan. Selftalk: A self-supervised commutative training diagram to comprehend 3d talking faces. In *Proceedings*

of the 31st ACM International Conference on Multimedia, pages 5292–5301, 2023.

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

[Ping *et al.*, 2013] Heng Yu Ping, Lili Nurliyana Abdullah, Puteri Suhaiza Sulaiman, and Alfian Abdul Halin. Computer facial animation: A review. *International Journal of Computer Theory and Engineering*, 5(4):658, 2013.

[Ren *et al.*, 2023] Zhiyuan Ren, Zhihong Pan, Xin Zhou, and Le Kang. Diffusion motion: Generate text-guided 3d human motion by diffusion model. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

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

[Shen and Tang, 2024] Fei Shen and Jinhui Tang. Imagpose: A unified conditional framework for pose-guided person generation. *Advances in neural information processing systems*, 37:6246–6266, 2024.

[Shen *et al.*, 2022] Shuai Shen, Wanhua Li, Zheng Zhu, Yueqi Duan, Jie Zhou, and Jiwen Lu. Learning dynamic facial radiance fields for few-shot talking head synthesis. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XII*, pages 666–682, 2022.

[Shen *et al.*, 2023] Shuai Shen, Wenliang Zhao, Zibin Meng, Wanhua Li, Zheng Zhu, Jie Zhou, and Jiwen Lu. Difftalk: Crafting diffusion models for generalized audio-driven portraits animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1982–1991, 2023.

[Shen *et al.*, 2025a] Fei Shen, Xin Jiang, Xin He, Hu Ye, Cong Wang, Xiaoyu Du, Zechao Li, and Jinhui Tang. Imagdressing-v1: Customizable virtual dressing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 6795–6804, 2025.

[Shen *et al.*, 2025b] Fei Shen, Cong Wang, Junyao Gao, Qin Guo, Jisheng Dang, Jinhui Tang, and Tat-Seng Chua. Long-term talkingface generation via motion-prior conditional diffusion model. *arXiv preprint arXiv:2502.09533*, 2025.

[Sohl-Dickstein *et al.*, 2015] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR, 2015.

[Song *et al.*, 2024] Wenfeng Song, Xuan Wang, Shi Zheng, Shuai Li, Aimin Hao, and Xia Hou. Talkingstyle: Personalized speech-driven 3d facial animation with style preservation. *IEEE Transactions on Visualization and Computer Graphics*, pages 1–12, 2024.

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

[Taylor *et al.*, 2012] Sarah L Taylor, Moshe Mahler, Barry-John Theobald, and Iain Matthews. Dynamic units of visual speech. In *Proceedings of the 11th ACM SIGGRAPH/Eurographics conference on Computer Animation*, pages 275–284, 2012.

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

[Wohlgenannt *et al.*, 2020] Isabell Wohlgenannt, Alexander Simons, and Stefan Stieglitz. Virtual reality. *Business & Information Systems Engineering*, 62:455–461, 2020.

[Wu *et al.*, 2023] Haozhe Wu, Songtao Zhou, Jia Jia, Junliang Xing, Qi Wen, and Xiang Wen. Speech-driven 3d face animation with composite and regional facial movements. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 6822–6830, 2023.

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

[Xu *et al.*, 2013] Yuyu Xu, Andrew W Feng, Stacy Marsella, and Ari Shapiro. A practical and configurable lip sync method for games. In *Proceedings of Motion on Games*, pages 131–140. 2013.

[Xu *et al.*, 2024] Sicheng Xu, Guojun Chen, Yu-Xiao Guo, Jiaolong Yang, Chong Li, Zhenyu Zang, Yizhong Zhang, Xin Tong, and Baining Guo. VASA-1: Lifelike audio-driven talking faces generated in real time. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[Zhang *et al.*, 2023] Chenxu Zhang, Chao Wang, Jianfeng Zhang, Hongyi Xu, Guoxian Song, You Xie, Linjie Luo, Yapeng Tian, Xiaohu Guo, and Jiashi Feng. Dream-talk: diffusion-based realistic emotional audio-driven method for single image talking face generation. *arXiv preprint arXiv:2312.13578*, 2023.