2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

3DFacePolicy: Speech-Driven 3D Facial Animation Based on Diffusion Policy

Xuanmeng Sha[1], Liyun Zhang[2*], Tomohiro Mashita[3], Naoya Chiba[1] and Yuki Uranishi[1]

***Abstract*— Speech-driven 3D facial animation has achieved significant progress in both research and applications. While recent baselines struggle to generate natural and continuous facial movements due to their frame-by-frame vertex generation approach, we propose 3DFacePolicy, a pioneer work that introduces a novel definition of vertex trajectory changes across consecutive frames through the concept of "action". By predicting action sequences for each vertex that encode frame-to-frame movements, we reformulate vertex generation approach into an action-based control paradigm. Specifically, we leverage a robotic control mechanism, diffusion policy, to predict action sequences conditioned on both audio and vertex states. Extensive experiments on VOCASET and BIWI datasets demonstrate that our approach significantly outperforms state-of-the-art methods and is particularly expert in dynamic, expressive and naturally smooth facial animations.**

I. INTRODUCTION

Speech-driven 3D facial animation creates realistic and precise 3D facial movements with vivid and natural expressions similar to real human on 3D vertex or blendshape templates with speech input. It is widely deployed in virtual digital human, AI assistant and digital twin robot learning [1], [2].

Recently, traditional generative methods can produce promising facial animations. Since the pioneering CNN-based approaches by [3], [4], the field has evolved significantly with Transformer-based architectures [5], [6], [7], [8]. However, these deterministic regression methods may lead to discontinuous facial animation due to the lack of explicit restraints with masking on discrete facial regions, thus overlooking realistic and natural human facial expressions.

The diffusion-based methods [9], [10] pioneer the integration of diffusion model to generate non-deterministic results with style conditions. However, these vertex-based generation methods may produce vague and discontinuous motions with high noise. Though 3DiFACE [11] employs vertex displacements, it overlooks the smoothness modeling on vertex movement trajectories, which leads to less natural facial animations.

*The corresponding author of this paper.
[1]Xuanmeng Sha, Naoya Chiba and Yuki Uranishi are with Graduate School of Information Science and Technology, The University of Osaka, Suita, Osaka 565-0871, Japan shaxuanmeng@gmail.com, chiba@nchiba.net, yuki.uranishi.cmc@osaka-u.ac.jp
[2]Liyun Zhang is with Graduate School of Frontier Sciences, The University of Tokyo, Kashiwa-shi, Chiba 277-8561, Japan liyun.zhang@lab.ime.cmc.osaka-u.ac.jp
[3]Tomohiro Mashita is with Department of Engineering Informatics, Osaka Electro-Communication University, Neyagawa, Osaka 572-8530, Japan mashita@osakac.ac.jp

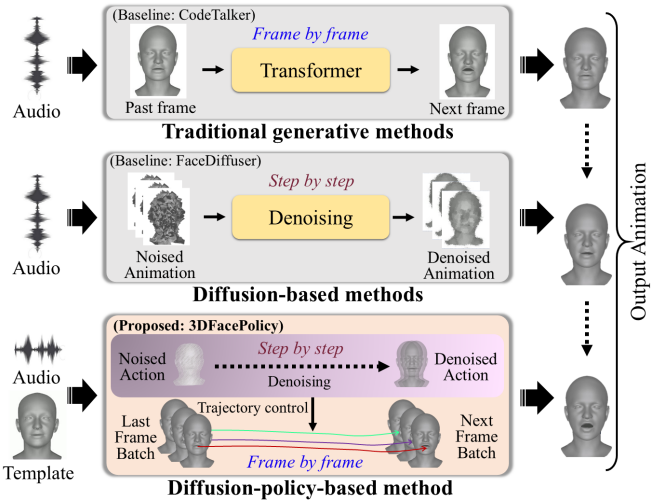

Fig. 1. We propose 3DFacePolicy, a 3D facial animation architecture that controls vertex movement trajectories through diffusion policy for action prediction. Unlike traditional generative methods like CodeTalker [6] that predict animation frame-by-frame using Transformer structures, or diffusion-based methods like FaceDiffuser [9] that generate animation from Gaussian noise step-by-step, our method reformulates vertex generation into trajectory control by accumulating denoised facial actions across consecutive frames through a robotic control mechanism.

To address these limitations, we propose 3DFacePolicy, a novel action-based facial vertex trajectory control model that generates smooth, continuous facial animation with realistic expressions and lip-sync accuracy based on a robotic-inspired diffusion policy framework. A conceptual comparison of our method with two other mainstream approaches is shown in Fig. 1. We reformulate the traditional vertex generation problem as a vertex trajectory control problem by innovatively defining "action" as temporal differential representations that encode kinematic variations between consecutive frames. This action space models both local temporal information and global spatial constraints, enabling more coherent motion synthesis compared to isolated frame-by-frame generation approaches.

For action prediction, we leverage diffusion policy [12], [13], a robot imitation learning framework that demonstrates high robustness on intensive and high-dimensional temporal representations. We adapt diffusion policy to predict vertex actions on 3D facial mesh, transforming facial animation synthesis from a vertex positioning problem into a motion trajectory prediction task. Through this action-based paradigm, predicted motion sequences are accumulated frame by frame to generate the final animation, naturally ensuring continu-

ous and smooth facial motions while maintaining realistic expressions and lip-sync accuracy.

The action sequences are first disentangled as temporal differential representations across consecutive frames, then facial movements are generated by sampling noisy action sequences from Gaussian noise and conditioning them on audio and vertex sequences using pretrained encoders, with final animation reconstructed by controlling vertex movement trajectories with the denoised action sequence. Extensive experiments demonstrate that our approach significantly outperforms state-of-the-art methods, ensuring smooth, flexible, and natural 3D facial animations.

The main contributions of our work are as follows:

- **A novel action-based control framework for 3D facial animation synthesis:** We design a pioneering framework that redefines 3D facial animation synthesis through an innovative action-based paradigm, introducing "action" to represent vertex trajectory changes across frames and providing a new baseline that transforms traditional vertex generation into motion trajectory control for more dynamic and natural facial motion synthesis.
- **An insight on vertex motion smoothness through control-based prediction:** Through extensive experiments, we discover that smoother vertex motion trajectories lead to more realistic and natural facial animations. Our state-of-the-art results and trajectory visualizations validate that motion continuity achieved through control-based methods is crucial for high-quality 3D facial animation synthesis, providing valuable guidance for future research.
- **A pioneer paradigm of treating 3D facial animation generation as facial vertex trajectory control:** We are the first to reformulate 3D facial animation synthesis as a vertex trajectory control problem by introducing robotic learning concepts. We adapt diffusion policy from robotics to treat facial dynamics as a motion control task, establishing a novel cross-domain paradigm that demonstrates the potential of applying robotic control principles to visual generation tasks.

II. Related Work

A. Speech-Driven 3D Facial Animation

Speech-driven 3D facial animation creates realistic and natural facial movements from speech, with crucial challenges in synchronizing tone, rhythm, and dynamics to mirror real human expressions. Research has mainly focused on traditional generative methods and diffusion-based methods.

1) Traditional generative methods: These methods design deterministic mappings from audio to facial motions using deep neural networks. Early works established foundational approaches but were limited to lip-only animations [14], [3], [15], [16], while later research expanded to full-face animations [4], [17]. Transformers [18] then emerged as a fundamental architecture, with works [5], [6] utilizing Wav2Vec 2.0 [19] for audio processing and VQ-VAE-inspired codebooks [20] for motion space representation. Recent works like UniTalker [7] introduce a unified multi-head architecture that can simultaneously output different 3D facial annotation types from a single audio input. KMTalk [8] handles cross-modal mapping uncertainty by embedding key motions before generating the full 3D sequence. Though these methods achieve promising results in lip synchronization and facial animation, discrete facial region processing may lead to discontinuous facial motion, and their deterministic architectures are limited in presenting dynamic facial movements [9].

2) Diffusion-based methods: For presenting diverse facial motions, the Denoising Diffusion Probabilistic Models [21], [22] guided by conditional data distributions are employed. In speech-driven 3D facial animation, FaceDiffuser [9] is the first to integrate the diffusion model into 3D facial animation synthesis. Furthermore, works [10], [11], [23], [24] focus on the head poses and personalized styles of speakers with conditional diffusion models. These methods present diversity on facial expression. Nevertheless, clear facial motions and compact contextual information are overlooked, which reduces the reality and consistency of human facial motions [10].

B. Diffusion Policy Models

The diffusion policy [12] is a visuomotor policy, which emerged as a crucial component in robotics [25] for enabling agents to perform complex tasks based on visual observations such as images or depth information. Recent approaches span various paradigms including reinforcement learning [26], [27], imitation learning [28], [29], and motion planning [30], [31]. Other works [32], [33] present multi-view condition or optimization process. 3D Diffusion Policy [13] presents a two-stage architecture combining perception and decision-making, achieving state-of-the-art performance in complex manipulation tasks. Meanwhile, image-to-image translation methods [34], [35], [36] have shown that capturing precise visual cues and scene-level information is essential for robust scene understanding.

Based on these methods, our work presents a pioneering paradigm that introduces diffusion policy to facial animation synthesis. The more dynamic, natural, and continuous facial motions with vertex trajectory control are generated based on this paradigm rather than deterministic vertex positioning in traditional generative methods and blurred movements in diffusion-based methods.

III. Methodology

A. Problem Formulation

In our method, we design facial movements diffusion policy model (3DFacePolicy) to control the trajectory of vertex movements in consecutive frames, represented as the action $a_0^{1:N} = (a_0^1, a_0^2, ..., a_0^N) \in \mathbb{R}^{N \times V \times 3}$, where $N, V, 3$ are the number of frames, mesh vertices, and dimensions. Conditioning on audio $s^{1:N}$ and vertices states $x^{1:N}$, the action $a_t^{1:N}$ from Gaussian Noise is gradually denoised into noise-free action sequence $a_0^{1:N}$, where $t \in \{1, ..., T\}$ is the

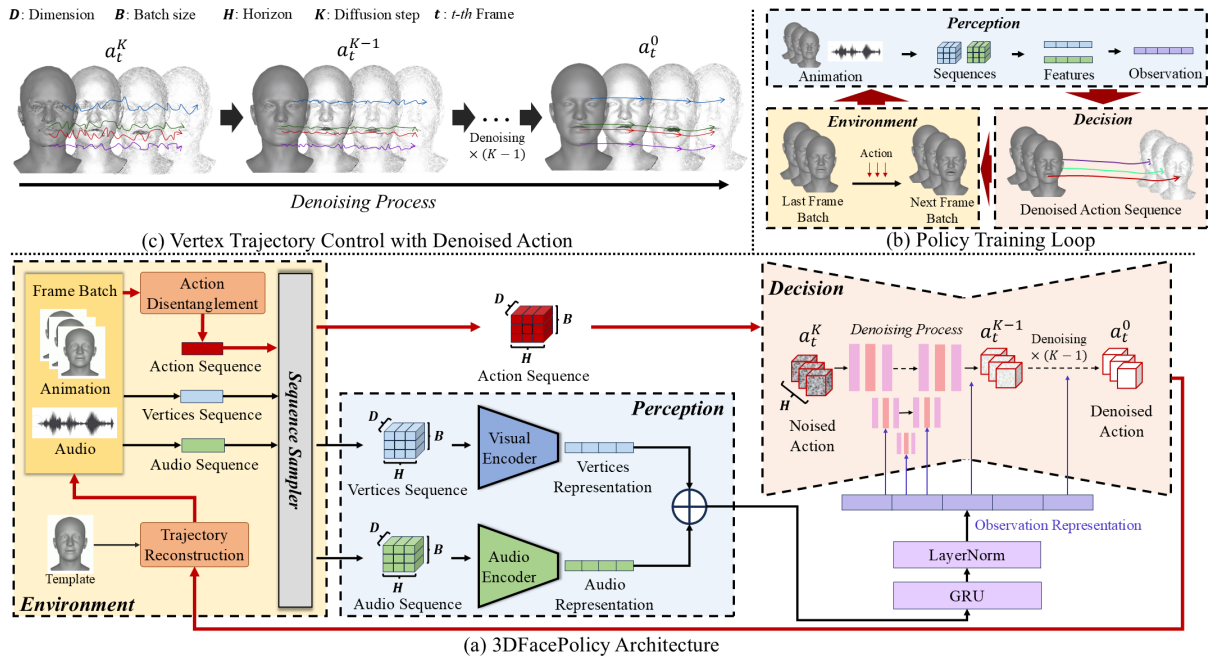

Fig. 2. **Overview of 3DFacePolicy architecture.** *(a)* Our architecture first disentangles animation into action sequences, then Perception module encodes vertices and audio sequences into observation representations that serve as conditions for Decision module, where actions are produced through denoising and control vertex trajectories on template to output animation. *(b)* The Environment, Perception, and Decision form a policy training loop for vertex trajectory control. *(c)* Noised action sequence sampled from Gaussian noise is gradually denoised into smooth and ordered actions for vertex trajectory control.

diffusion step. Therefore, our goal of the proposed architecture 3DFacePolicy is to control the movement trajectory of vertices with denoised action based on the conditional input of audio and vertices state. The problem could be formulated as:

$$a_0 = 3DFacePolicy(a_t, s, x, t), \tag{1}$$

With the predicted action sequence a_0 and the vertices of mesh template x_{temp}, the vertices in $n\text{-}th$ frame is presented as:

$$x_0^n = x_{temp} + \sum_{i=1}^{n} a_0^i, n \in \{1 : N\}, \tag{2}$$

where x_0^n is the $n\text{-}th$ frame of output animation $x_0^{1:N}$ with audio input following frame-by-frame predicted actions.

B. Architecture

1) Overview: We design our model with three modules following the policy training loop in robotics: Environment, Perception, and Decision, as illustrated in Fig. 2 (b). Environment module controls vertex trajectories with predicted action sequences and disentangles action sequences for next frame batches. Perception module generates comprehensive observation representation from vertices and audio sequences, serving as conditional input for guiding action policy learning process in Decision module. Decision module presents a conditioned denoising process (Fig. 2 (c)) where random and disordered action a_t^K are gradually denoised into temporally and spatially ordered action sequence a_t^0 conditioning on observations, ensuring smooth and natural motion trajectories for every vertex to generate dynamic and realistic facial animations. The overall architecture is shown in Fig. 2 (a).

2) Environment: The Environment module disentangles the action sequence from animation and samples the vertices, action, and audio sequence into a limited duration, ensuring the policy is trained in an action space with intensive context to maintain motion consistency and accuracy.

Action: A critical component of our approach is the formulation of action sequences that effectively capture facial motion dynamics. We define actions as temporal differential representations that encode the kinematic variations between consecutive frames with an adaptive scaling mechanism. This motion-centric formulation enables our model to learn smooth vertex movement trajectories.

Given a facial animation sequence with vertices $x^{1:N} = \{x^1, x^2, ..., x^N\} \in \mathbb{R}^{N \times V \times 3}$, the fundamental temporal displacement operator is expressed as:

$$\mathcal{D}_{temporal}^n = x^{n+1} - x^n, n \in [0, N-1] \tag{3}$$

This operator captures the raw inter-frame motion vectors that serve as the basis for action formulation.

Then we apply an adaptive scaling mechanism to enhance motion sensitivity and stability. An exponential weighting factor is leveraged to adaptively modulate the scaling based on the motion intensity. The scaling factor is defined as:

$$\Lambda_{adaptive} = \exp(-\beta \cdot \|\mathcal{D}_{temporal}^n\|_F^2) \tag{4}$$

The exponential function provides natural motion-based weighting where β controls the sensitivity decay rate, ensuring that subtle facial movements receive higher scaling

weights while preventing large motions. $\|\cdot\|_F$ denotes the Frobenius norm. Finally, the action is formulated as follows.

$$a^n = \varepsilon_{scaling} \cdot \Lambda_{adaptive} \cdot \mathcal{D}^n_{temporal} \tag{5}$$

where $\varepsilon_{scaling}$ is the base scaling parameter. The differential representation naturally preserves motion continuity through accumulation. Through empirical evaluation, this action-formulation-based method outperforms traditional vertex generation approaches.

After frame length alignment, vertices sequence $x^{1:N}$, audio sequence $s^{1:N}$, and action sequence $a^{1:N}$ with the same sequence length N are generated. Then sequence sampler samples data into manageable fixed durations called Horizon H, ensuring the action prediction policy is trained in a relatively local context with observation condition length N_{obs} and action-making length N_{act} to maintain smoothness and consistency. The sampled action sequence is isolated for coordination through the diffusion policy in the Decision module, while others serve as observation representations in the Perception module.

3) Perception: Perception module transforms the vertices and audio sequences with length H into a comprehensive representation $O = \{O_x, O_s\}$ that serves as conditions for Decision module, considered as a temporal observation fraction containing both audio and visual features. For visual data, we employ a lightweight encoder architecture inspired by [9], comprising linear layer, convolutional layer, and max-pooling layer to downsample the 3D features into a 1024-dimensional representation. The audio encoder utilizes the pretrained HuBERT model [37] to generate audio representations. The visual and audio features are concatenated and processed through a GRU layer followed by LayerNorm to generate comprehensive observation representations for Decision module.

4) Decision: The conditional denoising diffusion model is the backbone for learning facial action policy following [13]. For K iterations, A noised action sequence a_K is sampled from Gaussian noise with H length, conditioning on visual features x and audio features s, it is gradually denoised into a smooth and ordered action sequence a_0 with reverse process. The equation is formulated as follows:

$$a_{k-1} = \alpha_k(a_k - \gamma_k \epsilon_\theta(a_k, k, x, s)) + \sigma_k \mathcal{N}(0, \mathrm{I}), \tag{6}$$

where ϵ_θ is the denoising network, α_k, γ_k and σ_k are functions of k iteration. $\mathcal{N}(0, \mathrm{I})$ is Gaussian noise. After K iterations, the denoised action sequence is predicted.

For vertex trajectory reconstruction, with predicted action sequence $a_0^{1:N}$, the final vertex positions are computed as:

$$x_0^n = x_{temp} + \sum_{i=1}^{n} \mathcal{G}(a_0^i) \tag{7}$$

where x_{temp} is the neutral template mesh, $\mathcal{G}(\cdot)$ is the inverse transformation function, which presented as:

$$\mathcal{G}(a^i) = a^i / (\varepsilon_{scaling} \cdot \Lambda_{adaptive}) \tag{8}$$

5) Loss Function: In the diffusion process, the noise ϵ^θ is added on a randomly sampled action sequence a_0 at k iteration to train the denoising network ϵ_θ. The objective of this process is to predict the noise added on the sequence, which is defined as diffusion loss in our model:

$$\mathcal{L}_{\text{diff}} = \text{MSE}(\epsilon^k, \epsilon_\theta(\overline{\alpha_k} a_0 + \overline{\beta_k}\epsilon^k, k, x, s)), \tag{9}$$

where $\overline{\alpha_k}$ and $\overline{\beta_k}$ are noise schedules during diffusion steps. However, diffusion-loss-only training is insufficient for generating vertices sequences with smooth actions. Here, we also use reconstruction loss on vertex space to supervise the visual output:

$$\mathcal{L}_{\text{rec}} = \mathbb{E}_n[\frac{1}{N}\sum_{n=1}^{N} \|x_0^n - \hat{x}_0^n\|^2], \tag{10}$$

where x_0^n and $\hat{x}_0^n$ are the predicted vertices sequence x_0 and ground truth $\hat{x}_0$ in the $n\text{-}th$ frame from frame length N. We also use velocity loss to enhance the action smoothness:

$$\mathcal{L}_{\text{vel}} = \mathbb{E}_n[\frac{1}{N}\sum_{n=1}^{N} \|(x_0^{n-1} - x_0^n) - (\hat{x}_0^{n-1} - \hat{x}_0^n)\|^2]. \tag{11}$$

The total loss is the sum of these three losses:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{\text{diff}} + \lambda_2 \mathcal{L}_{\text{rec}} + \lambda_3 \mathcal{L}_{\text{vel}}. \tag{12}$$

Here, λ_1, λ_2, and λ_3 are weight coefficients that balance the relative importance of diffusion loss, reconstruction loss, and velocity loss in the total loss function.

6) Implementation details: We use DDIM [38] as denoising scheduler and sample prediction. For the sequence sampler, we set horizon length $H = 16$, $N_{obs} = N_{act} = 8$ and $\beta = 0.1$ in action scaling. Based on extensive experiments, the scaling product $\varepsilon_{scaling} \cdot \Lambda_{adaptive}$ exhibits consistent magnitudes within each dataset, and we accordingly apply 10^6 for VOCASET and 10^4 for BIWI to accommodate their respective motion intensity scales. For the trade-off parameters in loss function, λ_1, λ_2, λ_3 are empirically set to 1, 2, 0.5 respectively. We train for 600 epochs with batch size 32. The observation representation is 1024 dimensions. Our model is trained on a single V100 GPU with 32GB RAM. We employed the AdamW optimization algorithm, setting the learning rate to 0.0001 and gradually decreased to 0.000001.

IV. Experiments

We conduct comprehensive experiments to evaluate 3DFacePolicy on VOCASET [4] and BIWI [39]. Our evaluation includes both qualitative and quantitative analysis. We also conduct vertex trajectory smoothness comparison to further evaluate the contribution of our method. User study is also conducted as a convincing measurement to evaluate our method based on audiovisual user experience. Several ablation studies are also designed to analyze the impact of key settings in our model, specifically examining the effects of action definition, diffusion policy, horizon length, and loss function choices. Additional experimental results, architecture details, model computation efficiency analysis, and inference time are provided in the supplementary material [40].

TABLE I

QUANTITATIVE COMPARISON OF 3DFACEPOLICY WITH STATE-OF-THE-ART METHODS ON VOCASET AND BIWI DATASETS. OUR METHOD ACHIEVES OVERALL SUPERIOR PERFORMANCE COMPARED WITH OTHER BASELINES.

Method	VOCASET			BIWI		
	MVE ↓ $(10^{-3}mm)$	FDD ↓ $(10^{-7}mm)$	UFVE ↓ $(10^{-3}mm)$	MVE ↓ $(10^{-3}mm)$	FDD ↓ $(10^{-5}mm)$	UFVE ↓ $(10^{-3}mm)$
VOCA [4]	0.983	2.662	-	8.361	7.532	-
FaceFormer [5]	0.935	2.163	0.497	7.275	4.006	6.908
CodeTalker [6]	0.888	2.258	0.471	7.378	4.215	7.005
FaceDiffuser [9]	0.901	2.437	0.477	6.809	3.910	6.543
UniTalker [7]	0.853	2.390	0.449	**6.417**	5.044	6.148
3DFacePolicy	**0.847**	**1.502**	**0.416**	7.167	**1.778**	**5.433**

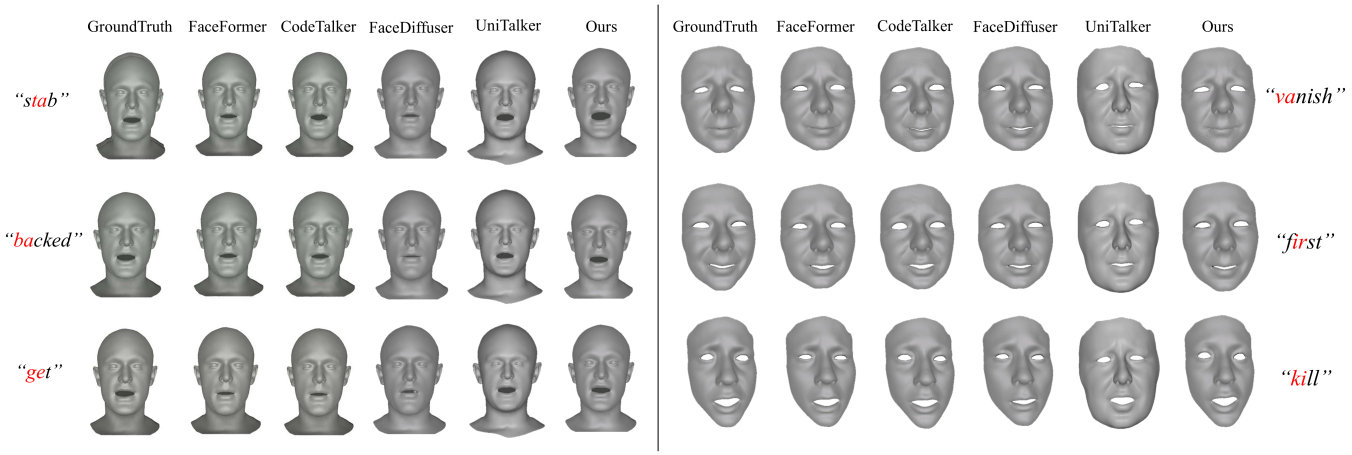

Fig. 3. Qualitative comparison of facial animation results on VOCASET (left) and BIWI (right). The figure shows the results of six different phonemes compared with other state-of-the-art results.

A. Experimental Settings

1) Baseline: We compare 3DFacePolicy with state-of-the-art methods: VOCA [4], FaceFormer [5], CodeTalker [6], FaceDiffuser [9] and UniTalker [7]. We focus on methods that emphasize vertex positioning accuracy and temporal consistency, as our primary contribution centers on trajectory smoothness rather than style controllability as in DiffPoseTalk [10]. For fair comparison, we use their official implementations and follow the same training/testing splits of datasets. All methods are evaluated under identical experimental settings to ensure valid comparisons.

2) Dataset: VOCASET dataset. [4] Contains 480 3D facial animation sequences with facial motions and audio from 12 subjects, recorded at 60 frames per second with 3-4 seconds duration. The template mesh uses FLAME [41] topology with 5023 vertices. We use the same training set (VOCA-Train), validation set (VOCA-Val), and test set (VOCA-Test) as in [6], [5]. *BIWI dataset.* [39] Contains recordings of 40 English sentences from 14 subjects, each read twice (neutral and emotional), captured at 25 frames per second with an average duration of 4.67 seconds. The 3D mesh template contains 23370 vertices. We use only the emotional sequences and follow the same dataset split as in [23], [5].

3) Evaluation Metric: To comprehensively evaluate the quality of generated facial animations, we employ Mean Vertex Error (MVE), Facial Dynamics Deviation (FDD) and Upper-face Vertex Error (UFVE) as our evaluation metrics following [7]. the MVE and UFVE measure the deviation of all face vertices and upper-face vertices respectively, while FDD measures the variation of facial dynamics for a motion sequence in comparison with ground truth.

B. Quantitative Evaluation

We conduct comprehensive quantitative evaluations on VOCASET and BIWI datasets, comparing our method with state-of-the-art approaches. Table I presents the comparison results using Mean Vertex Error (MVE), Facial Dynamics Deviation (FDD) and Upper-face Vertex Error (UFVE) metrics.

For the VOCASET dataset, our method achieves the best performance in all metrics with a notable reduction in FDD. For the BIWI dataset, we achieve the best FDD and UFVE scores, though with a slightly higher MVE score. This trade-off demonstrates our method's emphasis on capturing dynamic facial movements over strict vertex positioning accuracy. These results demonstrate 3DFacePolicy's effectiveness in modeling facial dynamics with motion trajectory control across datasets, validating our goal of producing smoother and more expressive facial animations while preserving temporal consistency and motion naturalness.

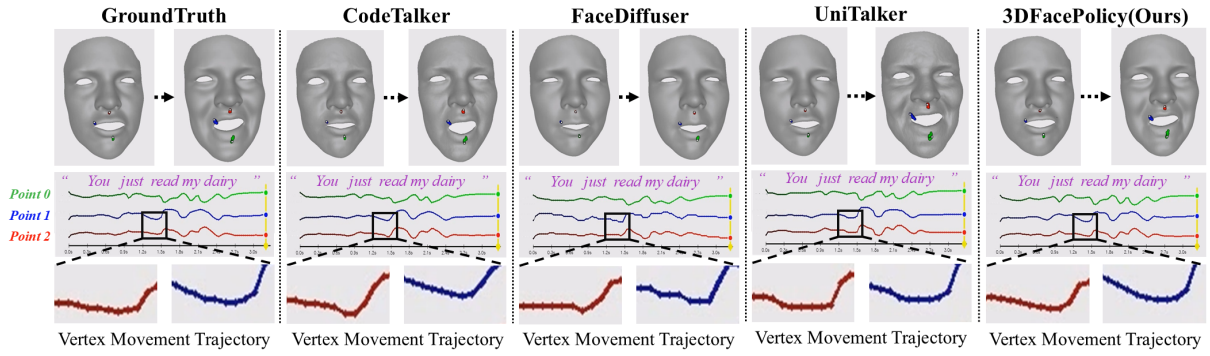

Fig. 4. Vertex movement trajectory comparison on BIWI dataset. Three randomly chosen vertex trajectories (shown in three different colors in the figure) with high motion amplitude in lip region are printed to compare the animation smoothness of 3DFacePolicy with other baselines.

TABLE II

USER STUDY RESULTS. PREFERENCE PERCENTAGE OF A/B TESTING ON LIP SYNC, REALISM AND EMOTIONAL EXPRESSION IS SHOWN.

Methods	Lip Sync (%) ↑		Realism (%) ↑		Emotional Expression (%) ↑	
	Ours	Competitor	Ours	Competitor	Ours	Competitor
Ours *vs.* CodeTalker [6]	**53.85**	46.15	**56.92**	43.08	**67.69**	32.31
Ours *vs.* FaceDiffuser [9]	**63.08**	36.92	**78.46**	21.54	**84.62**	15.38
Ours *vs.* UniTalker [7]	**57.78**	42.22	**73.33**	26.67	**66.67**	33.33
Ours *vs.* GT	41.54	**58.46**	**52.31**	47.69	**52.31**	47.69

C. Qualitative Evaluation

We visually evaluate our method against state-of-the-art methods including FaceFormer [5], CodeTalker [6], FaceDiffuser [9] and UniTalker [7]. The results are rendered on FLAME template of VOCASET and BIWI shown in Fig. 3.

Our method demonstrates natural and expressive facial movements with more realistic lip shapes during speech. When pronouncing syllables like "ki" in "kill" and "ba" in "backed", our method produces more realistic results than other methods, with mouth movements closely matching ground truth. Moreover, for expressive syllables like "va" in "vanish" and "ir" in "first", our method shows clearer emotions while maintaining natural mouth shapes than other baselines. The diverse facial movements and variable expressions are more explicit than existing methods, indicating that our approach prioritizes synthesizing dynamic expressions and realistic facial movements with vertex trajectory control over strict positioning.

D. Vertex Motion Smoothness Evaluation

We conduct experiments to visually evaluate the smoothness of facial vertex trajectories on generated animations over time with other state-of-the-art methods on BIWI dataset, randomly selecting three vertices in the lip region where motion amplitude is typically large and visualizing their movement trajectories.

By observing facial motion trajectories during high-amplitude segments such as the "re" sound in "read", our method produces significantly smoother transitions while other baselines exhibit uneven variations and noticeable abrupt shifts. These less smooth trajectories from vertex generation-based methods lead to subtle jittering in facial animations, with accumulation of such artifacts resulting in less natural and realistic outcomes. Our trajectory control-based approach effectively addresses these issues by controlling facial vertex movements that better approximate real human facial motion, demonstrating that smoother vertex trajectories directly contribute to more natural facial animation quality.

E. User Study

We conduct a user study to evaluate the quality of generated 3D talking faces following [7], selecting CodeTalker [6], FaceDiffuser [9], UniTalker [7] and ground truth as competitors in an A/B test. Participants watch randomly arranged paired videos of 3DFacePolicy and other works, evaluating three key metrics: lip synchronization, realism, and emotional expression. We randomly sampled 30 examples from VOCASET and BIWI with 15 video pairs per participant, resulting in 450 effective evaluation entries from 30 participants with good visual and auditory quality. Table II shows the statistical comparison of user preferences. Our method is preferred in all three metrics, greatly outperforming other state-of-the-art methods in realism and emotional expression while maintaining competitive performance with ground truth. For lip synchronization, 3DFacePolicy shows better preference than other baselines and slightly worse than ground truth as expected. Overall, 3DFacePolicy demonstrates explicit and realistic facial expressions, proving more natural and suitable for 3D facial animation through our facial motion control approach.

F. Ablation Study

We evaluate the effect of key elements in our proposed model: *(i) Action Definition.* Evaluating the impact of our

TABLE III

ABLATION STUDIES ON ACTION DEFINITION CHOICE, DIFFUSION POLICY, HORIZON LENGTH, AND LOSS FUNCTION COMPONENTS.

Action Definition Choice	MVE ↓	FDD ↓	UFVE ↓
w/o Action	5.278	11.977	4.913
w/o Adaptive Scaling	0.971	2.582	0.568
Diffusion Policy			
w/o Diffusion Policy	1.344	5.771	0.908
Horizon, Observation, Action Steps			
8, 4, 4	1.207	4.260	0.653
24, 12, 12	0.972	2.504	0.637
Loss Function			
w/o $\mathcal{L}_{\text{rec}}$	1.074	1.786	0.885
w/o $\mathcal{L}_{\text{vel}}$	0.887	1.705	0.434
w/o $\mathcal{L}_{\text{diff}}$	0.930	2.587	0.604
Full Model	**0.847**	**1.502**	**0.416**

key contribution on 3D facial animation synthesis; *(ii) Diffusion Policy.* Evaluating the robotic controlling mechanism; *(iii) Horizon Length.* Evaluating the impact of horizon length choice; *(iv) Loss Function.* Evaluating individual contributions of loss function components to model performance. Each experiment trains on VOCASET with MVE, FDD and UFVE units of $(\times 10^{-3} mm)$, $(\times 10^{-7} mm)$ and $(\times 10^{-3} mm)$ respectively. Results are shown in Table III.

1) Verification for Action definition choice: To verify the effectiveness of our proposed vertex action definition, we perform an ablation study by removing the scaling factor and whole action disentanglement module, which predicts actions without adaptive scaling and directly predicts facial animation in vertex space without action sequences.

Results show that removing action disentanglement significantly degrades performance across all metrics, highlighting its critical role in our framework. By separating actions from animation, the model can effectively control vertex trajectories conditioned on both audio and vertex states with smooth action sequences. Without action, the model struggles to establish clear relationships between audio input and facial movements, resulting in less accurate and natural facial animations, confirming that action is essential for generating accurate, smooth, and expressive facial animations and validates the effectiveness of reformulating generation into a controlling problem in 3D facial animation.

2) Verification for Diffusion Policy: To evaluate the effect of diffusion policy in 3D facial animation synthesis, we design a plain diffusion method without policy component that directly predicts entire action sequences with full animation length using only diffusion model, isolating the policy loop component's contribution and compare with other diffusion-based methods at action space level.

Results show that our diffusion policy-based method comprehensively outperforms the plain diffusion model. It demonstrates that directly inferring entire action sequences without the policy component lacks intensive contextual information, leading to unstable vertex trajectory control. It confirms that the robotic control methodology provides a stable and efficient paradigm for visual generation tasks such as 3D facial animation by transforming the problem from vertex generation to trajectory control, better addressing the smoothness of generated vertex trajectories and naturalness of facial motions.

3) Verification for Horizon Length: The horizon length determines the temporal context window for action prediction in facial motion synthesis. We evaluated three different settings of horizon H, observation condition length N_{obs} and action-making length N_{act}, while N_{obs} equals N_{act} and is half the length of horizon H. Horizon of 8 frames leads to less accurate predictions due to insufficient temporal context, while a longer horizon (24 frames) shows insufficient performance by overlooking intensive context information. The horizon length of 16 frames achieves optimal performance, particularly in capturing dynamic facial movements. Therefore, we set $H = 16$ as our default configuration.

4) Verification for Loss Function: In this experiment, we remove reconstruction loss $\mathcal{L}_{\text{rec}}$, velocity loss $\mathcal{L}_{\text{vel}}$ and diffusion loss $\mathcal{L}_{\text{diff}}$ respectively to assess the impact of each loss. Without $\mathcal{L}_{\text{rec}}$, both MVE and UFVE show surprising increases, indicating its role in maintaining vertex accuracy. Removing $\mathcal{L}_{\text{vel}}$ primarily affects FDD, suggesting its importance for temporal consistency. The absence of $\mathcal{L}_{\text{diff}}$ leads to a substantial increase in FDD and also affects MVE and UFVE, demonstrating its crucial part in maintaining smooth facial motions and model stabilization. Each component contributes meaningfully to the model's ability to predict dynamic, natural, and accurate facial animations.

V. CONCLUSION

In this paper, we propose 3DFacePolicy, a pioneering approach that reformulates speech-driven 3D facial animation from vertex generation to action-based trajectory control by introducing "action" as temporal differential representations encoding frame-to-frame vertex movements. We are the first to adapt diffusion policy from robotics to predict smooth action sequences conditioned on audio and vertex states, naturally ensuring smooth and realistic facial motions with accurate lip synchronization. Comprehensive experiments on VOCASET and BIWI datasets demonstrate that 3DFacePolicy significantly outperforms state-of-the-art approaches, particularly validating our insight that smoother vertex movement trajectories directly contribute to more natural animations. Future work will focus on adaptive sequence sampling strategies to enhance model flexibility while maintaining motion naturalness.

ACKNOWLEDGEMENT

This work was supported by JST BOOST, Japan Grant Number JPMJBS2402, and JSPS KAKENHI Grant Number JP21K14130, and Tateisi Science and Technology Foundation (Grant No. 2262008). We also acknowledge the use of Claude to refine this manuscript.

REFERENCES

[1] K. D. Yang, A. Ranjan, J.-H. R. Chang, R. Vemulapalli, and O. Tuzel, "Probabilistic speech-driven 3d facial motion synthesis: New benchmarks methods and applications," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 27 294–27 303.

[2] L. Zhang, Z. Luo, S. Wu, and Y. Nakashima, "Microemo: Time-sensitive multimodal emotion recognition with subtle clue dynamics in video dialogues," in *Proceedings of the 2nd International Workshop on Multimodal and Responsible Affective Computing*, 2024, pp. 110–115.

[3] T. Karras, T. Aila, S. Laine, A. Herva, and J. Lehtinen, "Audio-driven facial animation by joint end-to-end learning of pose and emotion," *ACM Transactions on Graphics*, p. 1–12, Aug 2017. [Online]. Available: http://dx.doi.org/10.1145/3072959.3073658

[4] D. Cudeiro, T. Bolkart, C. Laidlaw, A. Ranjan, and M. J. Black, "Capture, learning, and synthesis of 3d speaking styles," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 10 101–10 111.

[5] Y. Fan, Z. Lin, J. Saito, W. Wang, and T. Komura, "Faceformer: Speech-driven 3d facial animation with transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 18 770–18 780.

[6] J. Xing, M. Xia, Y. Zhang, X. Cun, J. Wang, and T.-T. Wong, "Codetalker: Speech-driven 3d facial animation with discrete motion prior," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 12 780–12 790.

[7] X. Fan, J. Li, Z. Lin, W. Xiao, and L. Yang, "Unitalker: Scaling up audio-driven 3d facial animation through a unified model," *arXiv preprint arXiv:2408.00762*, 2024.

[8] Z. Xu, S. Gong, J. Tang, L. Liang, Y. Huang, H. Li, and S. Huang, "Kmtalk: Speech-driven 3d facial animation with key motion embedding," *arXiv preprint arXiv:2409.01113*, 2024.

[9] S. Stan, K. I. Haque, and Z. Yumak, "Facediffuser: Speech-driven 3d facial animation synthesis using diffusion," in *Proceedings of the 16th ACM SIGGRAPH Conference on Motion, Interaction and Games*, 2023, pp. 1–11.

[10] Z. Sun, T. Lv, S. Ye, M. Lin, J. Sheng, Y.-H. Wen, M. Yu, and Y.-j. Liu, "Diffposetalk: Speech-driven stylistic 3d facial animation and head pose generation via diffusion models," *ACM Transactions on Graphics (TOG)*, vol. 43, no. 4, pp. 1–9, 2024.

[11] B. Thambiraja, S. Aliakbarian, D. Cosker, and J. Thies, "3diface: Diffusion-based speech-driven 3d facial animation and editing," *arXiv preprint arXiv:2312.00870*, 2023.

[12] C. Chi, S. Feng, Y. Du, Z. Xu, E. Cousineau, B. Burchfiel, and S. Song, "Diffusion policy: Visuomotor policy learning via action diffusion," *arXiv preprint arXiv:2303.04137*, 2023.

[13] Y. Ze, G. Zhang, K. Zhang, C. Hu, M. Wang, and H. Xu, "3d diffusion policy: Generalizable visuomotor policy learning via simple 3d representations," in *ICRA 2024 Workshop on 3D Visual Representations for Robot Manipulation*, 2024.

[14] S. Taylor, T. Kim, Y. Yue, M. Mahler, J. Krahe, A. G. Rodriguez, J. Hodgins, and I. Matthews, "A deep learning approach for generalized speech animation," *ACM Transactions On Graphics (TOG)*, vol. 36, no. 4, pp. 1–11, 2017.

[15] P. Edwards, C. Landreth, E. Fiume, and K. Singh, "Jali: an animator-centric viseme model for expressive lip synchronization," *ACM Transactions on graphics (TOG)*, vol. 35, no. 4, pp. 1–11, 2016.

[16] Y. Zhou, Z. Xu, C. Landreth, E. Kalogerakis, S. Maji, and K. Singh, "Visemenet: Audio-driven animator-centric speech animation," *ACM Transactions on Graphics (TOG)*, vol. 37, no. 4, pp. 1–10, 2018.

[17] A. Richard, M. Zollhöfer, Y. Wen, F. De la Torre, and Y. Sheikh, "Meshtalk: 3d face animation from speech using cross-modality disentanglement," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 1173–1182.

[18] A. Vaswani, "Attention is all you need," *Advances in Neural Information Processing Systems*, 2017.

[19] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 2020.

[20] A. Van Den Oord, O. Vinyals *et al.*, "Neural discrete representation learning," *Advances in neural information processing systems*, vol. 30, 2017.

[21] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.

[22] F.-A. Croitoru, V. Hondru, R. T. Ionescu, and M. Shah, "Diffusion models in vision: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 9, pp. 10 850–10 869, 2023.

[23] Z. Ma, X. Zhu, G. Qi, C. Qian, Z. Zhang, and Z. Lei, "Diffspeaker: Speech-driven 3d facial animation with diffusion transformer," *arXiv preprint arXiv:2402.05712*, 2024.

[24] H. K. Kim, S. Lee, and H. G. Kim, "Memorytalker: Personalized speech-driven 3d facial animation via audio-guided stylization," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025, pp. 11 241–11 251.

[25] X. Sha, Z. Jia, W. Sun, Y. Hao, X. Xiao, and H. Hu, "Development of mixed reality robot control system based on hololens," in *Intelligent Robotics and Applications*, H. Yu, J. Liu, L. Liu, Z. Ju, Y. Liu, and D. Zhou, Eds. Cham: Springer International Publishing, 2019, pp. 571–581.

[26] S. Levine, C. Finn, T. Darrell, and P. Abbeel, "End-to-end training of deep visuomotor policies," *Journal of Machine Learning Research*, vol. 17, no. 39, pp. 1–40, 2016.

[27] J. Lee, J. Hwangbo, L. Wellhausen, V. Koltun, and M. Hutter, "Learning quadrupedal locomotion over challenging terrain," *Science robotics*, vol. 5, no. 47, p. eabc5986, 2020.

[28] A. Rajeswaran, V. Kumar, A. Gupta, G. Vezzani, J. Schulman, E. Todorov, and S. Levine, "Learning complex dexterous manipulation with deep reinforcement learning and demonstrations," *arXiv preprint arXiv:1709.10087*, 2017.

[29] A. Zeng, P. Florence, J. Tompson, S. Welker, J. Chien, M. Attarian, T. Armstrong, I. Krasin, D. Duong, V. Sindhwani, and J. Lee, "Transporter networks: Rearranging the visual world for robotic manipulation," *Conference on Robot Learning (CoRL)*, 2020.

[30] B. Ichter, J. Harrison, and M. Pavone, "Learning sampling distributions for robot motion planning," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2018, pp. 7087–7094.

[31] P. Florence, C. Lynch, A. Zeng, O. Ramirez, A. Wahid, L. Downs, A. Wong, J. Lee, I. Mordatch, and J. Tompson, "Implicit behavioral cloning," *Conference on Robot Learning (CoRL)*, 2021.

[32] T.-W. Ke, N. Gkanatsios, and K. Fragkiadaki, "3d diffuser actor: Policy diffusion with 3d scene representations," *arXiv preprint arXiv:2402.10885*, 2024.

[33] A. Z. Ren, J. Lidard, L. L. Ankile, A. Simeonov, P. Agrawal, A. Majumdar, B. Burchfiel, H. Dai, and M. Simchowitz, "Diffusion policy policy optimization," in *arXiv preprint arXiv:2409.00588*, 2024.

[34] L. Zhang, P. Ratsamee, Y. Uranishi, M. Higashida, and H. Takemura, "Thermal-to-color image translation for enhancing visual odometry of thermal vision," in *2022 IEEE International Symposium on Safety, Security, and Rescue Robotics (SSRR)*. IEEE, 2022, pp. 33–40.

[35] L. Zhang, P. Ratsamee, B. Wang, Z. Luo, Y. Uranishi, M. Higashida, and H. Takemura, "Panoptic-aware image-to-image translation," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2023, pp. 259–268.

[36] L. Zhang, P. Ratsamee, Z. Luo, Y. Uranishi, M. Higashida, and H. Takemura, "Panoptic-level image-to-image translation for object recognition and visual odometry enhancement," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 2, pp. 938–954, 2023.

[37] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 29, pp. 3451–3460, 2021.

[38] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," *arXiv preprint arXiv:2010.02502*, 2020.

[39] G. Fanelli, J. Gall, H. Romsdorfer, T. Weise, and L. Van Gool, "A 3-d audio-visual corpus of affective communication," *IEEE Transactions on Multimedia*, vol. 12, no. 6, pp. 591–598, 2010.

[40] X. Sha, L. Zhang, T. Mashita, N. Chiba, and Y. Uranishi, "3dfacepolicy: Audio-driven 3d facial animation based on action control," 2025. [Online]. Available: https://arxiv.org/abs/2409.10848

[41] T. Li, T. Bolkart, M. J. Black, H. Li, and J. Romero, "Learning a model of facial shape and expression from 4d scans." *ACM Trans. Graph.*, vol. 36, no. 6, pp. 194–1, 2017.