MemoryTalker: Personalized Speech-Driven 3D Facial Animation via Audio-Guided Stylization

Hyung Kyu Kim[1] Sangmin Lee[2,*] Hak Gu Kim[3,*]

[1]Department of Imaging Science and Arts, Chung-Ang University, South Korea
[2]Department of Computer Science and Engineering, Korea University, South Korea
[3]Department of Metaverse Convergence, Chung-Ang University, South Korea

Abstract

Speech-driven 3D facial animation aims to synthesize realistic facial motion sequences from given audio, matching the speaker's speaking style. However, previous works often require priors such as class labels of a speaker or additional 3D facial meshes at inference, which makes them fail to reflect the speaking style and limits their practical use. To address these issues, we propose MemoryTalker which enables realistic and accurate 3D facial motion synthesis by reflecting speaking style only with audio input to maximize usability in applications. Our framework consists of two training stages: <1-stage> is storing and retrieving general motion (i.e., Memorizing), and <2-stage> is to perform the personalized facial motion synthesis (i.e., Animating) with the motion memory stylized by the audio-driven speaking style feature. In this second stage, our model learns about which facial motion types should be emphasized for a particular piece of audio. As a result, our MemoryTalker can generate a reliable personalized facial animation without additional prior information. With quantitative and qualitative evaluations, as well as user study, we show the effectiveness of our model and its performance enhancement for personalized facial animation over state-of-the-art methods. Project page: https://cau-irislab.github.io/ICCV25-MemoryTalker/

1. Introduction

Speech-driven 3D facial animation is a challenging task that aims to synthesize realistic 3D facial motion synchronized with the given speech [2, 5, 10, 11, 16, 20, 27, 31, 39, 46, 51, 53]. This technique can be widely utilized in various immersive applications, including VR telepresence, character animation for film production and gaming. Recently, the advent of large-scale 3D facial animation datasets [6, 13, 26, 34, 36] and advancements in deep learning have significantly increased interest in this field.

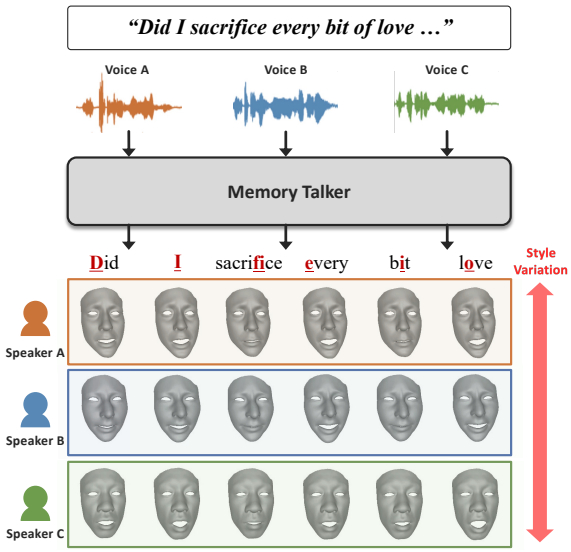


Figure 1. The intuition of *MemoryTalker* for personalized speech-driven 3D facial animation, *Memorizing* and *Animating*. **Memorizing**: Storing and retrieving facial motion. **Animating**: Synthesizing the personalized 3D facial motion with the stylized motion memory. Our MemoryTalker can accurately produce the personalized 3D facial motion for different speakers using only audio input.

The core challenge lies in developing algorithms that not only achieve precise speech-motion synchronization but also capture individual speaking styles - the subtle yet distinctive patterns in how different speakers articulate the same words.

This work addresses *personalization* issues in speech-driven 3D facial animation by considering the *speaking style* of a speaker, including the amplitude of mouth opening and closing, the extent of pouting, etc (see Fig. 1). To deal with that, most previous works [6, 12, 41, 45] have employed one-hot encoding of the identity classes for different speakers in the training set. However, these models necessarily require human identity classes even at inference time (Fig. 2 (a)). Furthermore, the inherent limitation of one-hot encoding

*Corresponding Author

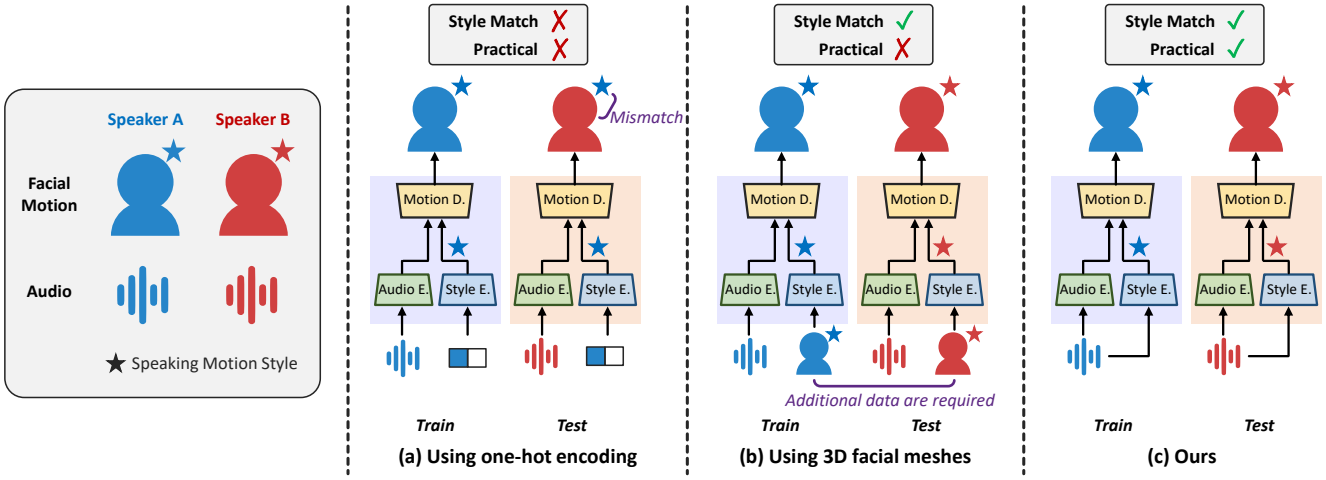


Figure 2. Explored approaches for personalized speech-driven 3D facial animation. (a) In the one-hot encoding approaches, while the identities of training speakers can be encoded with one-hot vector, it is impossible to match the unseen speaker's speaking style during inference. (b) In the approaches of utilizing 3D facial mesh sequence, by providing additional sequence of 3D facial mesh deformation during inference, they can produce the personalized 3D facial motion. However, it is not practical. (c) In our model, both during training and inference, personalized 3D facial animation is generated solely from the given audio input.

makes it impossible for them to deal with unseen speakers at all. The most recent studies [19, 44, 47] have explored using a sequence of 3D facial mesh deformation to control the different speaking styles of each speaker. Unlike one-hot encoding approaches, these methods can capture speaking styles of speakers by encoding the speaking style feature from their arbitrary facial motion sequences without class labels. Feeding 3D facial motion data would ideally reflect the speaking style, but requiring such additional data at inference is not practical for real-world applications (Fig. 2 (b)).

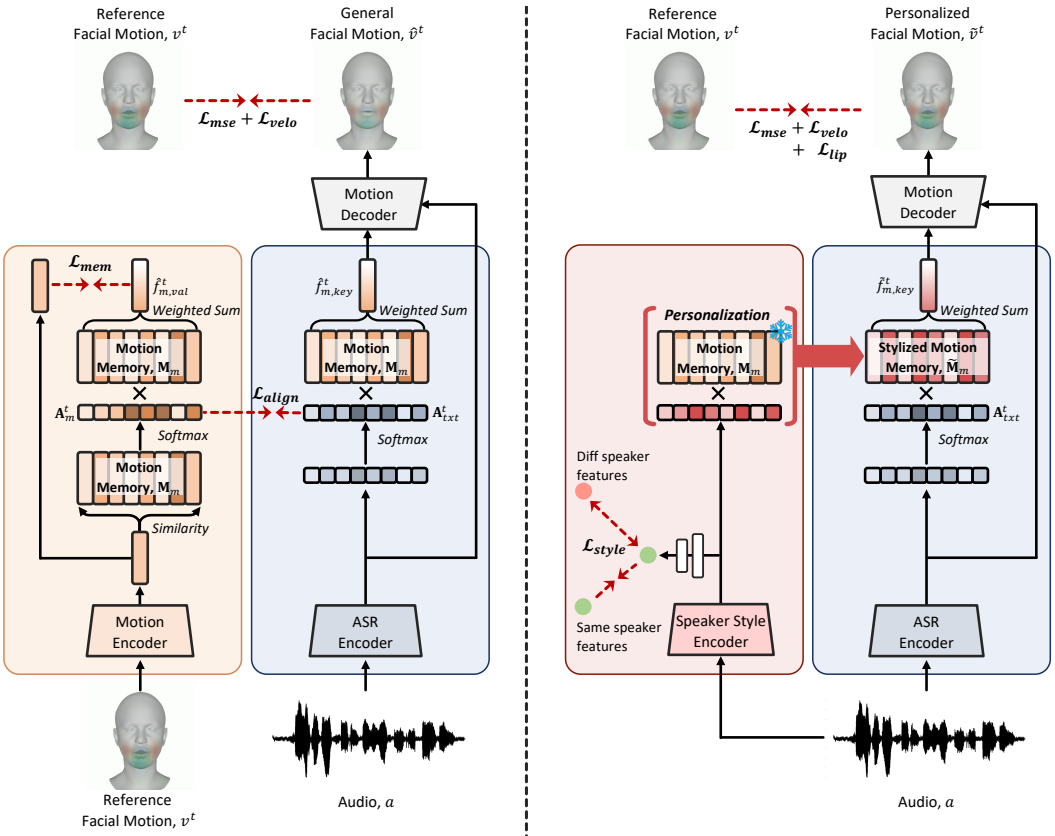
To overcome these limitations, we propose a novel speech-driven 3D facial animation model named *MemoryTalker* which can capture and predict speaking motion styles solely from audio input (see Fig. 2 (c)). It involves key-value multimodal memories that can effectively bridge different modalities. Our method consists of two training stages: <1-stage> is for storing general motion into a motion memory (*i.e.*, Memorizing), and <2-stage> is for stylizing the motion memory to synthesize personalized animations (*i.e.*, Animating).

In the first stage, we leverage a pre-trained ASR model to extract general motion feature representations with respect to the text aspect, ensuring consistent movements for a single phoneme. For example, when people say the word "who", the lips generally first come together and move forward to form the 'W' sound, then round to produce the 'OO' vowel sound. To explicitly store the facial motion feature, we design a motion memory and access the stored motion features using the text representations with a key-value structure. This key-value memory allows the model to map one modality to another effectively through accessing the stored motion features with different modal features in completely separate feature spaces. It alleviates the domain gap from inconsistent distributions of different modalities. However, it is difficult to synthesize accurate facial motion only with text representations because there exist different speaking styles even for the same word "who" such as variations in the extent of pouting.

To address this issue, in the second stage, our model is guided to generate personalized facial motions via audio-guided stylization. To this end, we personalize the trained motion memory based on audio signals. By distinguishing audio style features according to speaker types, we can achieve distinct style representations and refine the memory to synthesize the desired personalized motion effectively. Importantly, the proposed method does not require any prior knowledge (*e.g.*, ID class, additional facial meshes) at inference time, which makes our model more practical for real-world applications.

Our contributions are summarized as follows:

- We propose MemoryTalker, a novel framework for speech-driven 3D facial animation that can reflect speaking styles from audio input alone. To the best of our knowledge, this is the first work to address the personalization issue in this task without requiring additional prior information at inference time, which makes our approach practical.
- We introduce a new two-stage training strategy: (*i*) memorizing general facial motions aligned with neutral text representations, and (*ii*) animating personalized facial motions based on the stylized motion memory. These allow generating realistic animations that are accurate in terms of text content, while reflecting speaking styles effectively.



(a) 1-Stage: Storing and Recalling Facial Motion **(b) 2-Stage: Generating Personalized Facial Motion**

Figure 3. Illustration of the proposed *MemoryTalker* model for personalized 3D facial animation. (a) Learning to store facial motion feature in the facial motion memory and align motion features with text features. (b) Learning to disentangle unique speaking style of each speaker from audio and stylizing facial motion memory.

2. Related Work

2.1. Speech-Driven 3D Facial Animation

Earlier methods [1, 7, 9] for 3D facial animation usually employ a predefined facial template consisting of 3DMM parameters. Recently, various studies have been conducted to consider not only the mapping of speech to 3D facial mesh movements but also to reflect identity-specific speaking styles. Existing works [6, 12, 36, 45] have proposed a one-hot encoding of the identities of the training set. However, these models have a limitation in that they fail to predict new speaking styles during inference due to the use of the one-hot label slots for the subjects used in training.

To reflect subtle speaking styles better, Imitator [41] introduces a two-stage style adaptation to synthesize speaker-independent movement in the first stage and capture speaker-dependent style in the second stage using reference 2D video. To achieve careful control over speech-independent factors (e.g., speaking style), several methods have been proposed using reference 3D motion. Wu et al. [44] proposed an adaptive modulating module to consider both speech-independent composite and facial-dependent regional movements. Mimic [19] learned two disentangled latent spaces (content and style) for style-content disentanglement and 3D facial animation generation with identity-specific speaking styles. Finally, when generating 3D animation, they generate motion aligned to the audio in content space and apply a randomly provided speaking style in style space. Yang et al. [47] proposed a method to apply fine-detailed styles incrementally using reference 3D movements. Initially, coarse facial motion is learned from audio. Then, 3D facial motion with style is gradually learned from the provided reference 3D movement. However, these methods require a reference 2D video or a sequence of 3D facial meshes as well as audio signals as inputs during inference, which are resource-consuming and impractical for real-world applications. In contrast, we propose a novel approach that leverages speech-driven personalized memory networks to capture identity-specific speaking styles from speech signals alone.

2.2. Memory Network

Memory networks can enhance inference capability by being read from and written with external long-term memory components [28, 43]. The key-value memory network

architecture, enabling models to use keys to access relevant memories and retrieve corresponding values, has been widely adopted in various computer vision tasks including object tracking [14, 48], few-shot learning [4, 52], and anomaly detection [15, 23, 29, 30]. In recent years, the memory networks have been used in multi-modal modeling [21, 22, 24, 35, 50]. In [32, 40, 49], memory networks have been proposed to solve the alignment of associations between audio and visual information for 2D talking face generation. Compared to the previous works, we introduce a novel two-stage training strategy that enables to store and retrieve general motion and to predict personalized facial animation with the motion memory stylized by the audio-driven speaking style feature.

3. Proposed Method

Fig. 3 shows the overall framework of our personalized speech-driven 3D facial animation. To synthesize realistic 3D facial animation, we propose a two-stage training strategy: 1) storing and recalling general facial motion and 2) generating personalized facial motion with the stylized motion memory. In the first stage, our goal is to store facial motion features f_m^t in the motion memory $\mathbf{M}_m = \{s_m^i\}_{i=1}^n$ and recall general facial motion information from input audio a^t at time t. To achieve this, we encode the text representation f_{txt}^t from the audio signal and align it with f_m^t. This allows us to map the facial motion of various speakers for a single phoneme to the consistent text representation f_{txt}^t, thereby obtaining the general facial motion from the audio query a^t. In the second stage, we synthesize the personalized facial motion reflecting the speakers' speaking styles from audio $a^{1:T}$. To this end, we refine the pre-trained motion memory $\mathbf{M}_m$ into the stylized motion memory $\tilde{\mathbf{M}}_m$ while refining the text representation utilizing speaking style features f_s. The speaking style feature learns distinct characteristics from speaker-specific audio by adopting triplet loss. Finally, the recalled personal motion features from the stylized motion memory $\tilde{\mathbf{M}}_m$ are combined with the refined text representation and fed to the motion decoder to synthesize personalized 3D facial animation.

3.1. Facial Motion Memory

First, we design a motion memory $\mathbf{M}_m \in \mathbb{R}^{n\times c}$ with n slots and c channels to store motion feature $f_m^t \in \mathbb{R}^c$. The facial motion $v^t \in \mathbb{R}^3$ denotes the 3D movement of vertices over a neutral-face mesh template for each frame at time step t [45]. To this end, v^t is transformed into facial motion feature f_m^t using a motion encoder E_m. The encoded facial motion feature is used as a query to access our motion memory $\mathbf{M}_m$. When the facial motion feature f_m^t is given as a query in the motion memory $\mathbf{M}_m$, the attention weight w_m^i can be obtained by calculating the similarity between the f_m^t and each slot s_m^i in $\mathbf{M}_m$. This addressing procedure can be formulated as

$$w_m^i = \frac{\exp(\kappa \cdot d(s_m^i, f_m^t))}{\sum_{j=1}^n \exp(\kappa \cdot d(s_m^j, f_m^t))}, \tag{1}$$

where $d(\cdot,\cdot)$ indicates cosine similarity function. κ is a scaling factor. We omit the superscript t here for simplicity. Let $\mathbf{V}_m^t = \{w_m^1, w_m^2, \cdots, w_m^n\}$ denote the value address vector, which is a set of attention weights for the corresponding motion memory slots at time t. The recalled motion feature $\hat{f}_{m,val}^t$ can be retrieved using $\mathbf{V}_m^t$ and $\mathbf{M}_m$ by a weighted sum for each memory slot as follows:

$$\hat{f}_{m,val}^t = \sum_{i=1}^n w_m^i \cdot s_m^i. \tag{2}$$

To explicitly embed motion features into the motion memory, we adopt the memory reconstruction loss between the recalled and reference motion features.

$$\mathcal{L}_{mem} = \sum_{t=1}^T \left\| f_m^t - \hat{f}_{m,val}^t \right\|_2^2. \tag{3}$$

Motion information is stored in $\mathbf{M}_m$ by minimizing $\mathcal{L}_{mem}$.

We search for motion features that are synchronized with the input audio after storing memory. At this time, in order to minimize individual speaking styles of effect by the audio, such as pitch and tone, (*i.e.,* mapping multiple speaking style attributes to a single phoneme when the same pronouncing), we use the encoded text representation f_{txt}^t from the encoder of automatic speech recognition (ASR) E_{aud} as a query to access memory. This leads us to query common facial motion (i.e., common lip shapes for the same word regardless of the speaker). The encoded text representation f_{txt}^t can be written as

$$f_{txt}^{1:T} = \psi_{\rightarrow n}(\text{Interp}(E_{aud}(a))) \in \mathbb{R}^{T\times n}, \tag{4}$$

where $\psi_{\rightarrow n}(\cdot)$ is a single linear layer that projects to fit the number of slots. a represents the input audio signal. The pre-trained ASR encoder in HuBERT [18] is used as E_{aud} to map the source audio segment a^t to the text representation f_{txt}^t. Interp$(\cdot)$ is a linear interpolation function to synchronize the motion features and the text representations as video fps.

Let $\mathbf{K}_{txt}^t = \{w_{txt}^1, w_{txt}^2, \cdots, w_{txt}^n\}$ denote the key address vector, which is a set of attention weights for the corresponding motion memory slots. The key address vector $\mathbf{K}_{txt}^t$ is obtained by applying a softmax function to the projected text representation f_{txt}^t. It can be written as

$$\mathbf{K}_{txt}^t = \text{softmax}(f_{txt}^t) \in \mathbb{R}^n. \tag{5}$$

Then, the key address vector $\mathbf{K}_{txt}^t$ derived from the text representation f_{txt}^t is used to recall the general facial motion information across various speakers. Finally, the motion feature $\hat{f}_{m,key}^t \in \mathbb{R}^c$ is retrieved from motion memory using

the text address aligned to the motion address for general facial motion synthesis.

$$\hat{f}^t_{m,key} = \sum_{i=1}^{n} w^i_{txt} \cdot s^i_m. \tag{6}$$

To recall the corresponding motion features from $\mathbf{M}_m$ using key address vector $\mathbf{K}^t_{txt}$, it is required to align the key address vector $\mathbf{K}^t_{txt}$ with the value address vector $\mathbf{V}^t_m$. For this purpose, we employ KL divergence between them as an alignment loss $\mathcal{L}_{align}$, which can be defined as

$$\mathcal{L}_{align} = \sum_{t=1}^{T} \mathbf{K}^t_{txt} \log \frac{\mathbf{K}^t_{txt}}{\mathbf{V}^t_m}. \tag{7}$$

To synthesize facial motion v^t, we employ a motion decoder $D_m(\cdot)$ based on transformer decoder structure as in [12]. The facial motion can be generated as follows:

$$\hat{v}^t = D_m([f^t_{txt}; \hat{f}^t_{m,key}], f^t_{txt}). \tag{8}$$

The motion decoder is trained to minimize the reconstruction loss between the synthesized 3D facial motion $\hat{v}^t$ and ground truth 3D facial motion v^t. It can be written as

$$\mathcal{L}_{mse} = \sum_{t=1}^{T} \left\| v^t - \hat{v}^t \right\|^2. \tag{9}$$

In addition, we introduce a velocity loss $\mathcal{L}_{vel}$ to address the issue of jittery output frames when using only reconstruction loss [33]. $\mathcal{L}_{vel}$ is defined as

$$\mathcal{L}_{vel} = \sum_{t=0}^{T-1} \left\| (v^{t+1} - v^t) - (\hat{v}^{t+1} - \hat{v}^t) \right\|^2. \tag{10}$$

The total loss at the first training stage is defined as

$$\mathcal{L}_{\text{1-stage}} = \mathcal{L}_{mse} + \mathcal{L}_{vel} + \lambda_1(\mathcal{L}_{mem} + \mathcal{L}_{align}), \tag{11}$$

where we set λ_1 to 0.01.

3.2. Stylized Motion Memory

To synthesize the personalized 3D facial motion, the speaking style feature f_s is encoded from audio a. Without loss of generality, we assume that the audio signals include inherently speaking styles such as volume, pitch, and speaking speed. These speaking styles lead to differences in mouth opening size or the extent of pouting in 3D facial animation even when speakers say the same word or sentence.

To encode the speaking style feature f_s from audio a, we use the mel-spectrogram, which provides a representation aligned with human auditory perception. The speaking style feature $f_s \in \mathbb{R}^c$ can be defined as

$$f_s = E_s(\phi^{a2m}(a)), \tag{12}$$

where E_s is a speaking style encoder and ϕ^{a2m} is a function that converts audio into a mel-spectrogram.

In particular, to encode more discriminating style features, we introduce a speaking style loss $\mathcal{L}_{style}$ using triplet loss. The speaking style loss is designed to learn a feature space where similar speaking styles are closer (i.e., minimize intra-class variation) to each other and different speaking styles are further apart (i.e., maximize inter-class variation).

$$\mathcal{L}_{style} = \max\left(\|f_s - f^p_s\|^2_2 - \|f_s - f^n_s\|^2_2 + l, 0\right), \tag{13}$$

where f^p_s and f^n_s are the speaking styles of positive samples (i.e., same speaker) and negative samples (i.e., different speakers), respectively. They are obtained through the same process as f_s. l indicates a margin.

With the speaking style feature f_s, the pre-trained motion memory $\mathbf{M}_m$ can be updated to the stylized motion memory $\tilde{\mathbf{M}}_m = \{\tilde{s}^i_m\}^n_{i=1} \in \mathbb{R}^{n \times c}$, which can be written as

$$\tilde{\mathbf{M}}_m = \{\tilde{w}^i_s \cdot s^i_m\}^n_{i=1}, \tag{14}$$

where $\tilde{\mathbf{M}}_m$ is obtained by multiplying the style weight $\tilde{w}^i_s$ with the i-th memory slot in $\mathbf{M}_m$.

To update each memory slot, we design the style weight $\tilde{w}_s = \left\{\tilde{w}^i_s\right\}^n_{i=1} \in \mathbb{R}^n$ as follows:

$$\tilde{w}_s = \text{sigmoid}(\psi'_{\to n}(f_s)) \cdot \psi_{\to 1}(f_s) \tag{15}$$

where $\psi'_{\to n}$ and $\psi_{\to 1}$ are single linear layers that project to fit a n-dimensional vector and a scalar value, respectively. The values obtained from $\psi'_{\to n}(f_s)$ are processed with the sigmoid function to score each slot, and then scaled by the result of $\psi_{\to 1}(f_s)$. By applying the $\tilde{w}_s$ to the motion memory $\mathbf{M}_m$, the stylized motion memory $\tilde{\mathbf{M}}_m$ can reflect different speaking styles for each speaker via audio only.

Finally, the personalized motion $\tilde{v}^t \in \mathbb{R}^3$ can be reconstructed by incorporating the personalized motion feature $\tilde{f}_{m,key}$ recalled in $\tilde{\mathbf{M}}_m$.

$$\tilde{v}^t = D_m([f^t_{txt}; \tilde{f}^t_{m,key}], f^t_{txt}), \tag{16}$$

where $\tilde{f}^t_{m,key} = \sum^n_{i=1} w^i_{txt} \cdot \tilde{s}^i_m$, which is computed in the same way as the weighted sum used to calculate $f^t_{m,key}$.

As a result, by explicitly embedding the speaking style features into the motion memory and recalling it, we can produce more realistic and personalized 3D facial animation.

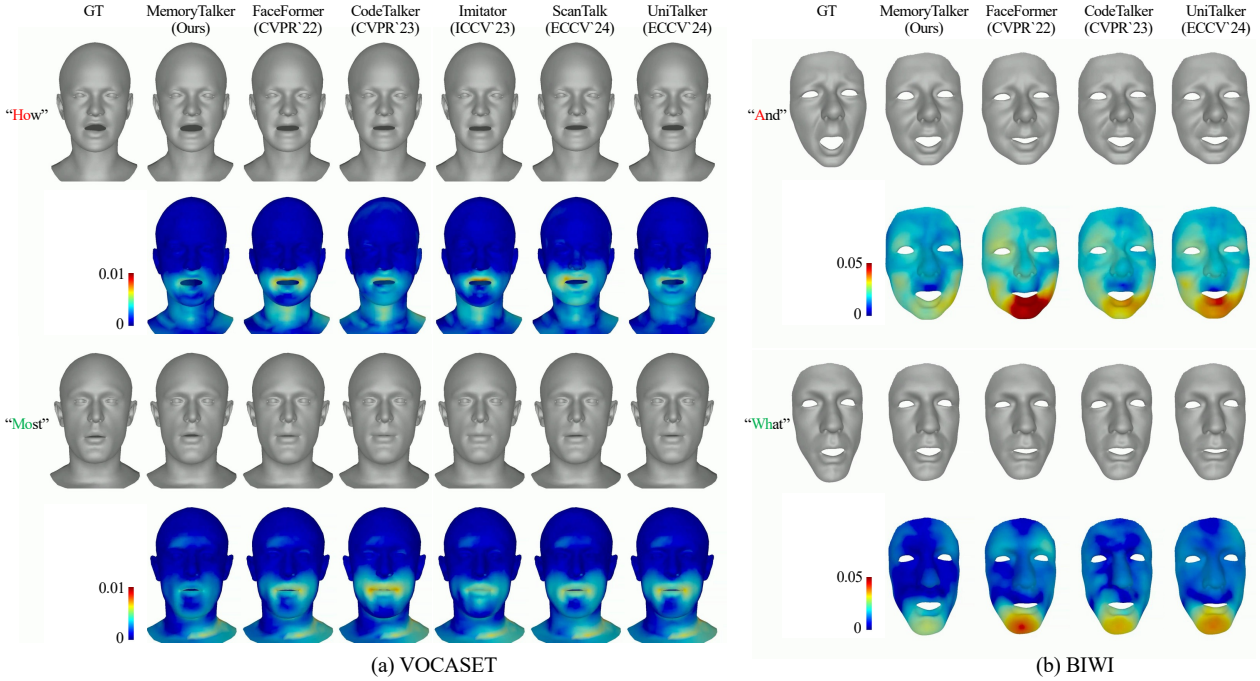
In the second stage, to train our model, we additionally employ lip vertex loss $\mathcal{L}_{lip}$, which is a mean square error for the lip region (i.e., lower-face vertices) to focus on carefully synthesizing lip movements according to various speaking styles. The total loss for the second stage is defined as

$$\mathcal{L}_{\text{2-stage}} = \mathcal{L}_{mse} + \mathcal{L}_{vel} + \lambda_2(\mathcal{L}_{lip} + \mathcal{L}_{style}), \tag{17}$$

where we set λ_2 to 0.01.

Method	VOCASET [6]					BIWI [13]				
	FVE ↓ ($\times 10^{-6}$)	LVE ↓ ($\times 10^{-5}$)	FID ↓ ($\times 10^{-1}$)	LDTW ↓ ($\times 10^{-5}$)	Lip-max ↓ ($\times 10^{-4}$)	FVE ↓ ($\times 10^{-4}$)	LVE ↓ ($\times 10^{-4}$)	FID ↓ ($\times 10^{-1}$)	LDTW ↓ ($\times 10^{-4}$)	Lip-max ↓ ($\times 10^{-3}$)
FaceFormer [12]	0.639	0.413	3.583	0.507	0.452	0.981	0.207	8.204	0.114	0.477
CodeTalker [45]	0.721	0.498	3.713	0.554	0.484	0.979	0.211	9.419	0.120	0.478
SelfTalk [33]	0.593	0.382	3.279	0.475	0.416	1.030	0.222	7.320	0.118	0.496
Imitator [41]	0.686	0.456	3.918	0.554	0.472	-	-	-	-	-
ScanTalk [31]	0.609	0.375	3.623	0.457	0.420	-	-	-	-	-
UniTalker [11]	0.570	0.382	3.256	0.507	0.407	0.919	0.196	7.234	0.109	0.461
MemoryTalker	**0.506**	**0.293**	**3.045**	**0.418**	**0.331**	**0.901**	**0.187**	**7.202**	**0.107**	**0.398**

Table 1. Quantitative evaluation for speech-driven 3D facial animation on VOCASET[6] and BIWI[13].



(a) VOCASET (b) BIWI

Figure 4. Visual comparisons with state-of-the art methods on (a) VOCASET and (b) BIWI. Note that the second and fourth rows represent the visualization of per-vertex errors.

4. Experiments

4.1. Datasets

We train and evaluate our model on VOCASET [6] and BIWI [13], which are widely used datasets for 3D facial animation. Both datasets contain audio-3D facial scan pairs that demonstrate English speech pronunciation. VOCASET [6] consists of 255 unique sentences, some shared across speakers, and contains 480 facial motion sequences from 12 subjects, captured at 60 fps for approximately 4 seconds per sequence. Each 3D face mesh is registered to the FLAME [25] topology with 5,023 vertices. BIWI [13] includes 40 unique sentences shared across all speakers. It consists of two parts: one with emotions and one without. There are 40 sentences uttered by 14 subjects. Each recording is repeated twice in neutral or emotional situations, capturing a dynamic 3D facial scan at 25 fps. The registered topology exhibits 23,370 vertices, with the average sequence length of 4.67s.

4.2. Implementation Details

Our experiments are conducted on an NVIDIA A6000 GPU. A single linear layer is used as motion encoder E_m as in [12, 45]. The structure in [8] is employed as our style encoder E_s. To encode the text representation from the audio, we utilize the pre-trained ASR model [18]. The projection layer $\psi_{\rightarrow n}$ in ASR consists of a single linear layer to project audio features onto the text logit and is fine-tuned during 1-stage training. In the first stage, we train our model (except for E_s) for 100 epochs with a learning rate of 0.0001. In the second stage, we freeze all layers of the first stage and train only the speaking style encoder E_s with a learning rate of 0.00005 for 100 epochs. The experimental protocol for both datasets

Method	FVE ↓ ($\times 10^{-6}$)	LVE ↓ ($\times 10^{-5}$)	FID ↓ ($\times 10^{-1}$)
FaceFormer [12]	$0.639^{\pm 0.036}$	$0.413^{\pm 0.058}$	$3.583^{\pm 0.358}$
CodeTalker [45]	$0.721^{\pm 0.056}$	$0.498^{\pm 0.037}$	$3.713^{\pm 0.373}$
Imitator [41]	$0.686^{\pm 0.069}$	$0.456^{\pm 0.067}$	$3.918^{\pm 0.536}$
MemoryTalker	**0.506**	**0.293**	**3.045**

Table 2. Error variability according to the used one-hot identity about a given audio on VOCASET [6].

Method	Inference Time	Parameter #
FaceFormer [12]	38.1 ms	92 M
CodeTalker [45]	297.6 ms	315 M
SelfTalk [33]	10.1 ms	450 M
UniTalker [11]	9.7 ms	313 M
MemoryTalker	7.8 ms	94 M

Table 3. Efficiency comparison of models based on inference time and number of parameters on VOCASET[6].

follows previous works [12, 45].

4.3. Evaluation Metrics

We adopt five quantitative evaluation metrics to evaluate the results: Face Vertex Error (FVE) and Lip Vertex Error (LVE) are the distance differences between the reference vertices and the generated vertices for the entire face and lip region, respectively. Lip Dynamic Time Warping (LDTW) is used to compute the temporal similarity of the lip region using DTW [37] as in [41]. Fréchet Inception Distance (FID) score [17] is used to evaluate the quality of the images rendered from the vertices [3]. In addition, we use Lip-max that averages the highest error among lip regions [36, 41].

4.4. Quantitative Results

Tab. 1 illustrates that our MemoryTalker outperforms state-of-the arts on both datasets. Our model achieves the lowest prediction errors. The lowest LDTW of our MemoryTalker means the highest temporal similarity for the lip region. Tab. 2 illustrates the problem of one-hot encoding approaches. They employ various identities in the training set for single audio during inference. Thus, they cannot reflect the speaking style of an unseen speaker during inference. As a result, the rendering quality for the same speaker can vary depending on the identity selected from the training set. On the other hand, our MemoryTalker can reflect the audio-driven speaking style feature, matching the speaker.

To evaluate the computational efficiency, as seen in Tab. 3, we measure the number of learnable parameters and inference time with a 1-second audio sample on VOCASET [6]. Compared to FaceFormer [12], CodeTalker [45], SelfTalk [33], and UniTalker [11], our MemoryTalker achieves

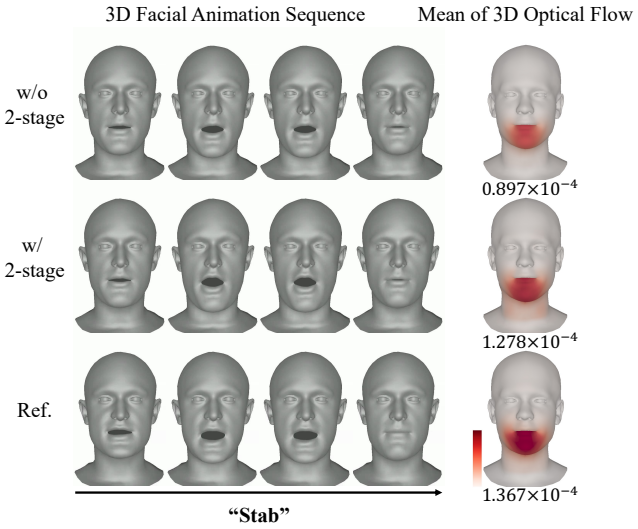


Figure 5. Qualitative results from "w/o 2-stage" and "w/ 2-stage". The last column represents the mean of 3D optical flow while pronouncing "**Stab**". The darker the red, the higher the motion.

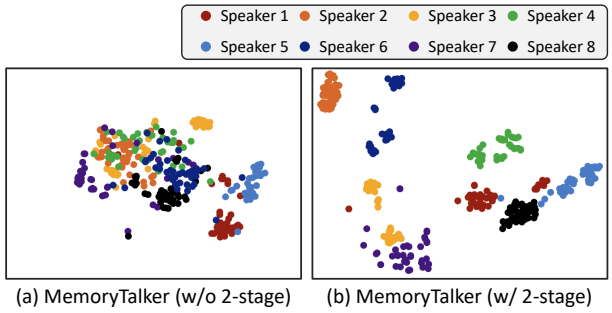


Figure 6. The t-SNE visualization of the recalled motion features across different speakers at each stage.

efficient 3D facial animations with about 120 fps.

4.5. Qualitative Results

4.5.1. Visual Comparisons

Fig. 4 shows the visual results of our MemoryTalker, compared to existing methods. When pronouncing "How" and "And" with a mouth open, our results provide more accurate mouth shapes. For pronunciations with mouth protrusion like "Most", "What", our results show similar mouth shapes to the reference (i.e., ground-truth). Although recent models generate some extent of mouth opening and closing according to the pronunciation, they lead to larger prediction errors around lip regions (lower-face vertices), compared to our MemoryTalker.

4.5.2. Adaptation of Speaking Style

To verify the effectiveness of our stylized motion memory at the 2-stage, we evaluate the dynamics of facial vertices

Proposed 1-stage training	Proposed 2-stage training	FVE ↓ ($\times 10^{-6}$)	LVE ↓ ($\times 10^{-5}$)
✗	✗	0.638	0.460
✓	✗	0.531	0.313
✓	✓	**0.506**	**0.293**

Table 4. Ablation study for various modules on VOCASET [6].

Loss	FVE ↓ ($\times 10^{-6}$)	LVE ↓ ($\times 10^{-5}$)	FID ↓ ($\times 10^{-1}$)	LDTW ↓ ($\times 10^{-5}$)
w/o $\mathcal{L}_{lip}$	0.514	0.297	3.057	0.435
w/o $\mathcal{L}_{style}$	0.513	0.295	3.058	0.431
Full	**0.506**	**0.293**	**3.045**	**0.418**

Table 5. Ablation study for our components on VOCASET [6].

Competitors	Lip Sync(%)	Realism(%)	Speaking Style(%)
vs. FaceFormer [12]	83.9	85.5	80.6
vs. CodeTalker [45]	85.5	83.9	71.8
vs. Imitator [41]	87.1	87.1	78.2
vs. ScanTalk [31]	94.4	91.1	92.8
vs. UniTalker [11]	79.8	80.6	86.3

Table 6. User study: our method vs. competitors on VOCASET [6].

sequence using 3D optical flows. Fig. 5 shows an example of visual results and the mean of optical flows from 1-stage and 2-stage when pronouncing "**Stab**". As shown in Fig. 5, the model trained at both stages delivers more accurate lip movements than the one trained at 1-stage only (without training the model at 2-stage).

To explore the effectiveness of the stylized motion memory in the latent space, we visualize the recalled features before decoding each subject's facial motion at each stage using t-SNE [42]. In Fig. 6, each point represents facial motion features for each sentence, and different colors indicate different speakers. Fig. 6 (a) shows that the distribution of each speaker is not separated because general facial motion is retrieved at 1-stage. In Fig. 6 (b), the distribution for each speaker is well-clustered, indicating that the speaking style of each speaker is captured at 2-stage.

4.6. Ablation Study

4.6.1. Memory and Style Encoder

To demonstrate the effectiveness of two-stage training, we evaluate the performance at each stage. In Tab. 4, the baseline (w/o 1-stage and 2-stage) is a model without our motion memory structure and is trained by adopting common reconstruction loss $\mathcal{L}_{mse}$ and velocity loss $\mathcal{L}_{vel}$. As seen in Tab 4, by explicitly storing and recalling the facial motion in 1-stage, FVE and LVE are reduced. It means that the facial motion recalled from the facial motion memory can support the facial synthesis results from text representation. In addition, by incorporating the personalized facial motion through the stylized motion memory in 2-stage, our model achieves the best performance.

4.6.2. Constraints

We investigate the impact of removing different constraints which are $\mathcal{L}_{lip}$ and $\mathcal{L}_{style}$ (see Tab. 5). When removing the $\mathcal{L}_{lip}$, performances are slightly worse than those with $\mathcal{L}_{lip}$. This is because the lip vertex loss is only relevant to the lower-face vertices. When removing the triplet loss for $\mathcal{L}_{style}$, we can observe that training becomes unstable and performance drops significantly. It highlights the importance of effectively learning personal speaking styles.

4.7. User Study

We conduct A/B tests on VOCASET to evaluate perceptual lip-sync, realism, and speaking style of synthesized 3D facial animations. A total of 33 subjects participated in the user study. To evaluate speaking style, lip-sync, and realism, we compare our MemoryTalker with 5 other methods (FaceFormer, CodeTalker, Imitator, ScanTalk, UniTalker) on 50 samples. To measure lip-sync and realism, we present each pair of identical sentences to the subject and the subject chooses the better one. To evaluate speaking style, we present participants with samples from our model, a comparison model, and the ground truth (reference). Participants are then asked to choose which of the two model-generated samples appears to have a speaking style most similar to the ground truth, considering factors such as the amplitude of mouth opening and closing, the degree of pouting, and other relevant aspects. As shown in Table 6, our results are generally favorable over other state-of-the-art methods across a variety of perceptual factors.

5. Conclusion

In this study, we demonstrate the ability of motion memory networks to bridge different modalities, which are 3D facial motion and speech, without requiring additional prior information for personalized speech-driven facial animations. To effectively train our MemoryTalker model, we introduce a two-stage training strategy: 1) storing and retrieving general motion and 2) performing the personalized 3D facial motion synthesis with the stylized motion memory. Extensive experimental results unequivocally highlight the superior performance of our MemoryTalker compared to existing speech-driven 3D facial animation models. In particular, our MemoryTalker achieves more favorable results not only in quantitative evaluations but also in the user study from a perceptual perspective. We hope this work can pave the way for practical personalized speech-driven 3D facial animation in real-world VR and metaverse applications.

6. Acknowledgments

This research was supported by the MSIT(Ministry of Science and ICT), Korea, under the Graduate School of Metaverse Convergence support program(IITP-2025-RS-2024-00418847) supervised by the IITP(Institute for Information & Communications Technology Planning & Evaluation) and the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (RS-2023-00253232)

References

[1] A. Verma, N. Rajput, and L.V. Subramaniam. Using viseme based acoustic models for speech driven lip synthesis. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03).*, pages V–720, 2003. 3

[2] Shivangi Aneja, Justus Thies, Angela Dai, and Matthias Nießner. Facetalk: Audio-driven motion diffusion for neural parametric head models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21263–21273, 2024. 1

[3] Elif Bozkurt. Personalized speech-driven expressive 3d facial animation synthesis with style control. *arXiv preprint arXiv:2310.17011*, 2023. 7

[4] Qi Cai, Yingwei Pan, Ting Yao, Chenggang Yan, and Tao Mei. Memory matching networks for one-shot image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4080–4088, 2018. 4

[5] Aggelina Chatziagapi and Dimitris Samaras. Avface: Towards detailed audio-visual 4d face reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16878–16889, 2023. 1

[6] Daniel Cudeiro, Timo Bolkart, Cassidy Laidlaw, Anurag Ranjan, and Michael Black. Capture, learning, and synthesis of 3D speaking styles. *Computer Vision and Pattern Recognition (CVPR)*, pages 10101–10111, 2019. 1, 3, 6, 7, 8

[7] José Mario De Martino, Léo Pini Magalhães, and Fábio Violaro. Facial animation based on context-dependent visemes. *Computers & Graphics*, 30(6):971–980, 2006. 3

[8] Zongyang Du, Berrak Sisman, Kun Zhou, and Haizhou Li. Disentanglement of emotional style and speaker identity for expressive voice conversion. In *Interspeech 2022*, pages 2603–2607, 2022. 6

[9] 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. 3

[10] Han EunGi, Oh Hyun-Bin, Kim Sung-Bin, Corentin Nivelet Etcheberry, Suekyeong Nam, Janghoon Ju, and Tae-Hyun Oh. Enhancing speech-driven 3d facial animation with audio-visual guidance from lip reading expert. In *Interspeech 2024*, pages 2940–2944, 2024. 1

[11] Xiangyu Fan, Jiaqi Li, Zhiqian Lin, Weiye Xiao, and Lei Yang. Unitalker: Scaling up audio-driven 3d facial animation through a unified model. In *European Conference on Computer Vision*, pages 204–221. Springer, 2024. 1, 6, 7, 8, 2

[12] 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 (CVPR)*, 2022. 1, 3, 5, 6, 7, 8, 2

[13] 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. 1, 6

[14] Zhihong Fu, Qingjie Liu, Zehua Fu, and Yunhong Wang. Stmtrack: Template-free visual tracking with space-time memory networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13774–13783, 2021. 4

[15] Dong Gong, Lingqiao Liu, Vuong Le, Budhaditya Saha, Moussa Reda Mansour, Svetha Venkatesh, and Anton van den Hengel. Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection. In *IEEE International Conference on Computer Vision (ICCV)*, 2019. 4

[16] Shan He, Haonan He, Shuo Yang, Xiaoyan Wu, Pengcheng Xia, Bing Yin, Cong Liu, Lirong Dai, and Chang Xu. Speech4mesh: Speech-assisted monocular 3d facial reconstruction for speech-driven 3d facial animation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14192–14202, 2023. 1

[17] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 7

[18] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460, 2021. 4, 6

[19] 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 *The 38th Annual AAAI*

Conference on Artificial Intelligence (AAAI), 2024. 2, 3, 1, 5, 6

[20] Hyung Kyu Kim, Sangmin Lee, and Hak Gu Kim. Analyzing visible articulatory movements in speech production for speech-driven 3d facial animation. In *2024 IEEE International Conference on Image Processing (ICIP)*, pages 3575–3579. IEEE, 2024. 1

[21] Minsu Kim, Joanna Hong, Se Jin Park, and Yong Man Ro. Multi-modality associative bridging through memory: Speech sound recollected from face video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 296–306, 2021. 4

[22] Minsu Kim, Jeong Hun Yeo, and Yong Man Ro. Distinguishing homophenes using multi-head visual-audio memory for lip reading. In *Proceedings of the AAAI conference on artificial intelligence*, pages 1174–1182, 2022. 4

[23] Sangmin Lee, Hak Gu Kim, Dae Hwi Choi, Hyung-Il Kim, and Yong Man Ro. Video prediction recalling long-term motion context via memory alignment learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. 4

[24] Sangmin Lee, Hyung-Il Kim, and Yong Man Ro. Weakly paired associative learning for sound and image representations via bimodal associative memory. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10534–10543, 2022. 4

[25] Tianye Li, Timo Bolkart, Michael J Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4d scans. *ACM Trans. Graph.*, 36(6): 194–1, 2017. 6

[26] Haiyang Liu, Zihao Zhu, Giorgio Becherini, Yichen Peng, Mingyang Su, You Zhou, Xuefei Zhe, Naoya Iwamoto, Bo Zheng, and Michael J Black. Emage: Towards unified holistic co-speech gesture generation via expressive masked audio gesture modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1144–1154, 2024. 1

[27] Jingying Liu, Binyuan Hui, Kun Li, Yunke Liu, Yu-Kun Lai, Yuxiang Zhang, Yebin Liu, and Jingyu Yang. Geometry-guided dense perspective network for speech-driven facial animation. *IEEE Transactions on Visualization and Computer Graphics*, 28(12):4873–4886, 2021. 1

[28] Alexander Miller, Adam Fisch, Jesse Dodge, Amir-Hossein Karimi, Antoine Bordes, and Jason Weston. Key-value memory networks for directly reading documents. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1400–1409, Austin, Texas, 2016. Association for Computational Linguistics. 3

[29] Byeongjun Min, Jihoon Yoo, Sangsoo Kim, Dongil Shin, and Dongkyoo Shin. Network anomaly detection using memory-augmented deep autoencoder. *IEEE Access*, 9:104695–104706, 2021. 4

[30] Chaoxi Niu, Guansong Pang, and Ling Chen. Graph-level anomaly detection via hierarchical memory networks. In *Machine Learning and Knowledge Discovery in Databases: Research Track: European Conference, ECML PKDD 2023, Turin, Italy, September 18–22, 2023, Proceedings, Part I*, page 201–218, Berlin, Heidelberg, 2023. Springer-Verlag. 4

[31] Federico Nocentini, Thomas Besnier, Claudio Ferrari, Sylvain Arguillere, Stefano Berretti, and Mohamed Daoudi. Scantalk: 3d talking heads from unregistered scans. In *European Conference on Computer Vision*, pages 19–36. Springer, 2024. 1, 6, 8

[32] 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*, pages 2062–2070, 2022. 4

[33] 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*, page 5292–5301, 2023. 5, 6, 7

[34] 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, 2023. 1

[35] Darshana Priyasad, Tharindu Fernando, Simon Denman, Sridha Sridharan, and Clinton Fookes. Memory based fusion for multi-modal deep learning. *Information Fusion*, 67:136–146, 2021. 4

[36] Alexander Richard, Michael Zollhöfer, Yandong Wen, Fernando de la Torre, and Yaser Sheikh. Meshtalk: 3d face animation from speech using cross-modality disentanglement. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1153–1162, 2021. 1, 3, 7

[37] Stan Salvador and Philip Chan. Toward accurate dynamic time warping in linear time and space. *Intelligent Data Analysis*, 11(5):561–580, 2007. 7

[38] RJ Skerry-Ryan, Eric Battenberg, Ying Xiao, Yuxuan Wang, Daisy Stanton, Joel Shor, Ron Weiss, Rob Clark, and Rif A Saurous. Towards end-to-end prosody transfer for expressive speech synthesis with tacotron. In *international conference on machine learning*, pages 4693–4702. PMLR, 2018. 2

[39] Zhiyao Sun, Tian Lv, Sheng Ye, Matthieu Lin, Jenny Sheng, Yu-Hui Wen, Minjing Yu, and Yong-jin Liu. Diffposetalk: Speech-driven stylistic 3d facial animation and head pose generation via diffusion models. *ACM Transactions on Graphics (TOG)*, 43(4):1–9, 2024. 1

[40] Shuai Tan, Bin Ji, and Ye Pan. Emmn: Emotional motion memory network for audio-driven emotional talking face generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22146–22156, 2023. 4

[41] Balamurugan Thambiraja, Ikhsanul Habibie, Sadegh Aliakbarian, Darren Cosker, Christian Theobalt, and Justus Thies. Imitator: Personalized speech-driven 3d facial animation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 20621–20631, 2023. 1, 3, 6, 7, 8, 5

[42] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008. 8, 2

[43] Jason Weston, Sumit Chopra, and Antoine Bordes. Memory networks. *arXiv preprint arXiv:1410.3916*, 2014. 3

[44] 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*, page 6822–6830, New York, NY, USA, 2023. Association for Computing Machinery. 2, 3, 1

[45] 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. 1, 3, 4, 6, 7, 8, 2, 5

[46] Zhihao Xu, Shengjie Gong, Jiapeng Tang, Lingyu Liang, Yining Huang, Haojie Li, and Shuangping Huang. Kmtalk: Speech-driven 3d facial animation with key motion embedding. In *European Conference on Computer Vision*, pages 236–253. Springer, 2024. 1

[47] Karren D Yang, Anurag Ranjan, Jen-Hao Rick Chang, Raviteja Vemulapalli, and Oncel 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*, pages 27294–27303, 2024. 2, 3

[48] Tianyu Yang and Antoni B. Chan. Learning Dynamic Memory Networks for Object Tracking. In *ECCV*, 2018. 4

[49] Ran Yi, Zipeng Ye, Juyong Zhang, Hujun Bao, and Yong-Jin Liu. Audio-driven talking face video generation with learning-based personalized head pose. *arXiv preprint arXiv:2002.10137*, 2020. 4

[50] Long Ying, Hui Yu, Jinguang Wang, Yongze Ji, and Shengsheng Qian. Fake news detection via multimodal topic memory network. *IEEE Access*, 9:132818–132829, 2021. 4

[51] Qingcheng Zhao, Pengyu Long, Qixuan Zhang, Dafei Qin, Han Liang, Longwen Zhang, Yingliang Zhang, Jingyi Yu, and Lan Xu. Media2face: Co-speech facial animation generation with multi-modality guidance. In *ACM SIGGRAPH 2024 Conference Papers*, New York, NY, USA, 2024. Association for Computing Machinery. 1

[52] Linchao Zhu and Yi Yang. Inflated episodic memory with region self-attention for long-tailed visual recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4344–4353, 2020. 4

[53] Yixiang Zhuang, Baoping Cheng, Yao Cheng, Yuntao Jin, Renshuai Liu, Chengyang Li, Xuan Cheng, Jing Liao, and Juncong Lin. Learn2talk: 3d talking face learns from 2d talking face. *IEEE Transactions on Visualization and Computer Graphics*, 2024. 1

MemoryTalker: Personalized Speech-Driven 3D Facial Animation via Audio-Guided Stylization

Supplementary Material

A. Overview

This supplementary document provides additional experiments and explanations to complement the main manuscript. We present a comprehensive performance comparison focusing on the effects of different memory slot sizes in our model. Additionally, we offer quantitative results comparing end-to-end training versus our proposed 2-stage training strategy. The document also includes extensive qualitative results, featuring detailed frame-by-frame comparisons between our method and existing approaches for different speakers under various conditions. Furthermore, we provide an in-depth feature analysis, visualizing how our method effectively captures speaking styles from audio input effectively. Lastly, we detail the implementation specifics of our MemoryTalker architecture and describe the methodology of our user study for a thorough evaluation against other methods.

B. Quantitative Results

B.1. Effects of Memory Slot Size

We conduct experiments varying the number of memory slots. Figure S1 shows Lip Vertex Errors (LVE) across different slot size configurations. We validate the performance while varying the memory slot size to (16, 24, 32, 48, 64). The optimal performance is achieved with 32 slots (indicated by star marker in Figure S1). Note that we leverage memory slot size 32 for our experiments in the main manuscript.

B.2. End-to-End Learning vs. Two-stage Training Strategy

We propose a novel two-stage training strategy that first induces general motion synthesis, followed by speaking style adaptation. Table S1 shows performance comparisons between end-to-end learning and our two-stage training strategy. As seen in the Table S1, our training strategy consistently outperforms the end-to-end learning across all evaluation metrics, including FVE, LVE, and FID. These results validate that our strategy more effectively captures and reproduces personalized speaking styles while maintaining motion accuracy.

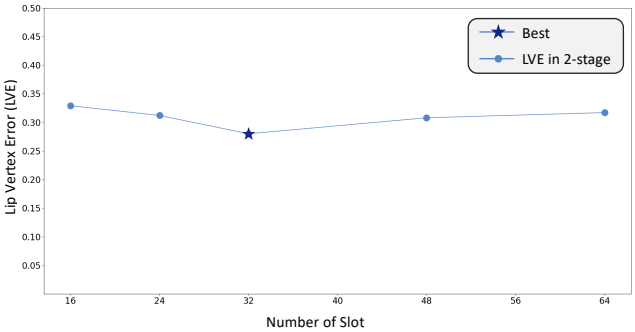


Figure S1. Effects of the memory slot size on LVE for facial animation generation.

Training Strategy	FVE ↓ ($\times 10^{-6}$)	LVE ↓ ($\times 10^{-5}$)	FID ↓ ($\times 10^{-1}$)
End-to-End	0.510	0.303	3.142
Two-Stage (Ours)	**0.506**	**0.293**	**3.045**

Table S1. Performance comparison between end-to-end learning and two-stage training strategy.

Method	FVE ↓ ($\times 10^{-6}$)	LVE ↓ ($\times 10^{-5}$)	LDTW ↓ ($\times 10^{-5}$)	Lip-max ↓ ($\times 10^{-4}$)
Mesh-driven [44]	0.673	0.400	0.521	0.431
Audio-driven (Ours)	**0.506**	**0.293**	**0.418**	**0.331**

Table S2. Quantitative results about mesh-based method.

Method	FVE↓ ($\times 10^{-1}$)	LVE↓ ($\times 10^{-3}$)	LDTW↓ ($\times 10^{-5}$)	Lip-max↓ ($\times 10^{-1}$)
CodeTalker [45]	0.122	0.453	0.637	0.356
UniTalker [11]	0.101	0.283	0.569	0.305
MemoryTalker (Ours)	**0.094**	**0.266**	**0.557**	**0.285**

Table S3. Quantitative results about seen identities.

B.3. Comparison with Mesh-based Methods

As presented in Table S2, we conduct a quantitative comparison with an additional mesh-based method [44]. The results clearly indicate that our audio-driven approach achieves superior performance across all evaluation metrics. It is worth noting that Mimic [19], another mesh-based method, could not be directly compared as it utilizes different datasets for training and evaluation.

Method	FVE ↓ ($\times 10^{-6}$)	LVE ↓ ($\times 10^{-5}$)	LDTW↓ ($\times 10^{-5}$)	Params (M)
ASR for E_s	0.518	0.300	0.437	185
MemoryTalker (Ours)	**0.506**	**0.293**	**0.418**	**94**

Table S4. Comparison the style features extracted by Mel-spectrogram- and ASR.

B.4. Evaluation on Seen Identities

We further evaluate our model's performance on seen identities, with the results shown in Table S3. The evaluation is performed on the BIWI Test-A set. We compare our **MemoryTalker** with CodeTalker [45], a one-hot-based method, and UniTalker [11], the current state-of-the-art approach. For a fair comparison, the correct one-hot identity vector is provided to CodeTalker. As demonstrated in the table, our method outperforms both competing methods in the seen identity setting as well.

B.5. Analysis on Speaking Style Features

We justify our choice of using mel-spectrograms for the speaking style encoder. Mel-spectrograms are well-known for preserving rich acoustic details such as prosody, rhythm, and timbre [38]. In contrast, features from an Automatic Speech Recognition (ASR) model are less suitable for this task because ASR models are trained to extract neutral text representations, thereby suppressing speaker-specific identity and style.

To empirically validate this, we present an ablation study in Table S4, where we replace our mel-spectrogram-based style encoder with an ASR-based one. The results show a degradation in performance, confirming that mel-spectrograms are better suited for capturing the nuances of speaking style for our task.

C. Qualitative Results

C.1. Performances Comparisons with Existing Methods

Figures S2 and S3 show extensive qualitative results, including detailed performance comparisons with existing methods. Our analysis provides a frame-by-frame visualization, offering a more nuanced and comprehensive comparison between our proposed method and current state-of-the-art approaches. Figure S2 illustrates cases where the pronunciation necessitates the opening of the mouth, exemplified by sounds like /a/. Conversely, Figure S3 showcases instances where the pronunciation involves the initial stages of lip closure, as demonstrated by sounds such as /o/. Upon careful examination of the error maps, it becomes evident that our proposed method achieves significantly more accurate results across both pronunciation types and diverse speaker profiles. This superior performance is consistently maintained when compared to all existing methods in the field. These results indicate that our method effectively captures and generates the respective speaking styles of multiple people under various different conditions.

C.2. The Effectiveness of Stylized Motion Memory in 2-Stage

Figure S6 shows the effectiveness of the stylized motion memory in 2-stage. In Fig. S6, the first row shows the facial mesh and error map rendered by the *general* motion feature retrieved from motion memory $\mathbf{M}_m$ in 1-stage. On the other hand, the fourth row shows the facial mesh and error map rendered by the *personalized* motion feature recalled from the stylized motion memory $\tilde{\mathbf{M}}_m$ in 2-stage. The second and third rows show the magnified versions of the lip regions of the first and fourth rows, respectively. As shown in Fig. S6, the limited lip movement can be observed in the first and second rows (results of 1-stage only, i.e., learning without 2-stage) compared to the third and fourth rows (results of learning with 2-stage). These results show that the motion memory $\mathbf{M}_m$ in 1-stage has a limitation in representing the subtle personal speaking style. On the other hand, by recalling the personalized motion feature from the stylized motion memory $\tilde{\mathbf{M}}_m$ in 2-stage, the 3D facial mesh can achieve fewer errors reflecting the individual's delicate speaking style.

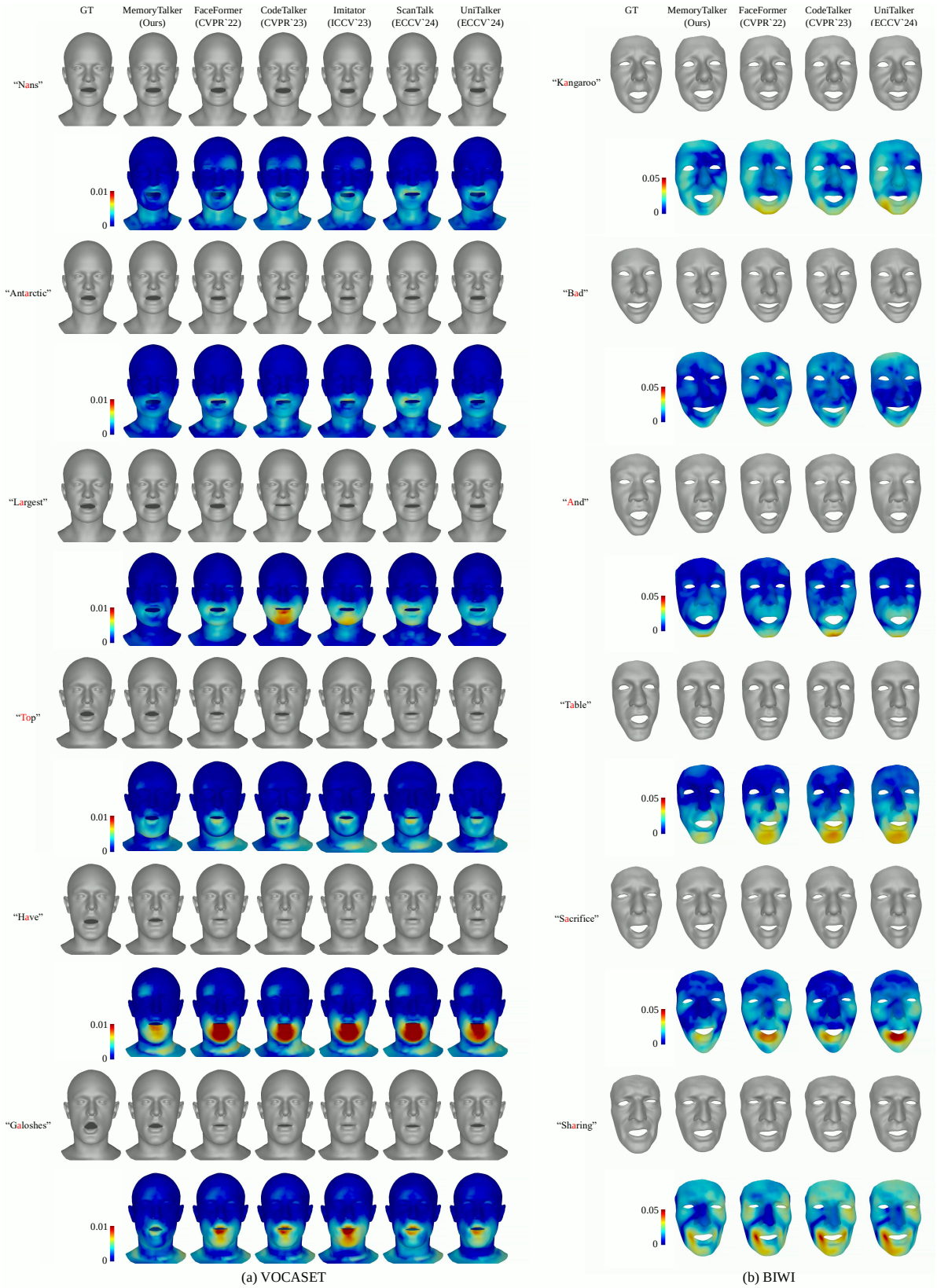
C.3. Applying Style feature to General Motion

Figure S4 demonstrates the effectiveness of reflecting the style feature in 2-stage. In Figure S4, the first row shows the generated 3D facial motion from synchronized with audio at 1-stage. At this time, general motion corresponding to the audio is synthesized. On the other hand, the second, thrid, fourth, and fifth row demonstrates when the style features generated differently for each person in the 2-stage were applied to the general motion generated in the 1-stage, the subtle changed lip shape in the 2-stage was visualized. Through this, we verify that our method not only supplements the information lacking in general motion through speaking style features, but also effectively learns speaking style information for each speaker.

D. Feature Analysis

D.1. Speaking Style Feature Visualization

We provide visualizations of motion features with t-SNE [42] to show that our method reflects speaking styles in the audio, contrasting it with existing approaches (see Figure S7). Figure S7 (a) visualizes the motion feature synthesized during inference when using one-hot encoding [12]. As in [12], since there are not able to know one-hot class informa-



(a) VOCASET (b) BIWI

Figure S2. Qualitative comparisons for the cases where the pronunciation involves starting to OPENING the mouth.

(a) VOCASET (b) BIWI

Figure S3. Qualitative comparisons for the cases where the pronunciation involves starting to CLOSING the lips.

Module	Input→Output	Operation
Motion Encoder	$\mathbf{Motion}(T, V \times 3) \rightarrow f_m(T, d_m)$	Linear(V × 3, d_m)
Speaker Style Encoder	$\mathbf{MelSpec}(T_{mel}, 80) \rightarrow \tilde{f}_s(d_{txt})$	Conv1d(80, 128) × 8 → [B, T_{mel}, 1280] Conv1d(1280, 128) → [B, T_{mel}, 128] GroupNorm(128/16, 128) Conv1d(128, 128) × 6 → [B, T_{mel}, 128] GroupNorm(128/16, 128) Conv1d(128, 128) × 6 → [B, T_{mel}/8, 128] GroupNorm(128/16, 128) AdaptiveAvgPool1d(1) → [B, 128] Linear(128, 128) × 6 Linear(128, 128) × 6 Linear(128, 256) → [B, 256] Linear(256, d_{txt}) → [B, d_{txt}]
Motion Memory	$f_{txt}(d_{txt}) \rightarrow f_m(d_m)$	Parameter(d_{txt}, d_m)
ASR Encoder	$\mathbf{Audio}(T_a) \rightarrow f_{txt}(T, d_{txt})$	Conv1d(1, d_{ASR}) → [B, T_a, d_{ASR}] LayerNorm(d_{ASR}) LinearInterpolation → [B, T, d_{ASR}] Conv1d(d_{ASR}, d_{ASR}) × 5 → [B, T, d_{ASR}] LayerNorm(d_{ASR}) Transformer(d_{ASR}) Linear(d_{ASR}, d_{txt}) → [B, T, d_{txt}]
Motion Decoder	$f_m(T, d_m + d_{txt}) \rightarrow \mathbf{Motion}(T, V \times 3)$	Linear($d_m + d_{txt}$, d_m) Transformer(d_m) Linear(d_m, $V \times 3$)

Table S5. The detailed architecture of our MemoryTalker.

tion, one-hot vectors are arbitrarily selected (8 classes). As a result, this model cannot distinguish actual unseen speakers at all. Figure S7 (b) shows the results of utilizing 3D facial mesh sequence [19]. It considers 3D facial mesh sequences as additional inputs. However, the 3D facial meshes are usually unavailable in real-world situations at inference. In addition, it does not distinguish the style distribution corresponding to the unseen speakers. On the other hand, as shown in Figure S7 (c), our model provides clear speakers' clustering that corresponds to individual speaking styles. In particular, the proposed method does not require any prior information (*i.e.*, speaker information) in both training and inference stages, which makes it more practical in real-world scenarios.

D.2. Key Addressing Vector Visualization

To verify which key addressing vector correctly retrieves the motion memory, we visualize the key addressing vectors of our model. Figure S5 shows the generated 3D facial mesh sequences and the corresponding key addressing vectors of MemoryTalker model. The key addressing vectors are generated from an audio segment pertaining to the phoneme "/a:/" and "/w/", respectively, for different speakers. We can see that the address smoothly varies as the lip region in the mesh moves. From visualizing the key addressing vectors of different speakers speaking the same pronunciation, we observe the similar tendency of the key addressing vectors. Focusing on the slots that noticeably change their address value, from the 3rd column, the address on the 8th and 16th slot addresses increases from 0.075 to 0.200 when pronouncing "/a:/", and also when presented with other speakers' speech. Similarly, when pronouncing "/w/", it shows that the address on the 2nd and 17th slot addresses increases regardless of speaker. These demonstrate that the key addressing vectors of our MemoryTalker are activated in the same slot when synthesizing the same lip shape, suggesting that our motion memory slot feature accurately stores the corresponding lip motion. Note that the key addressing vector here is from the ASR model, which is to extract speaker-neutral general motions. Our personalization is further applied to the memory components, not the key addressing vector.

E. Implementation Details

E.1. Details of MemoryTalker Architecture

We configured the main models as in Table S5. This is the baseline architecture used in our all experiments. The Motion Encoder is designed as a single linear layer as in [12, 19, 41, 45]. First, the Speaking Style Encoder concatenates the output and input features of each layer to create a

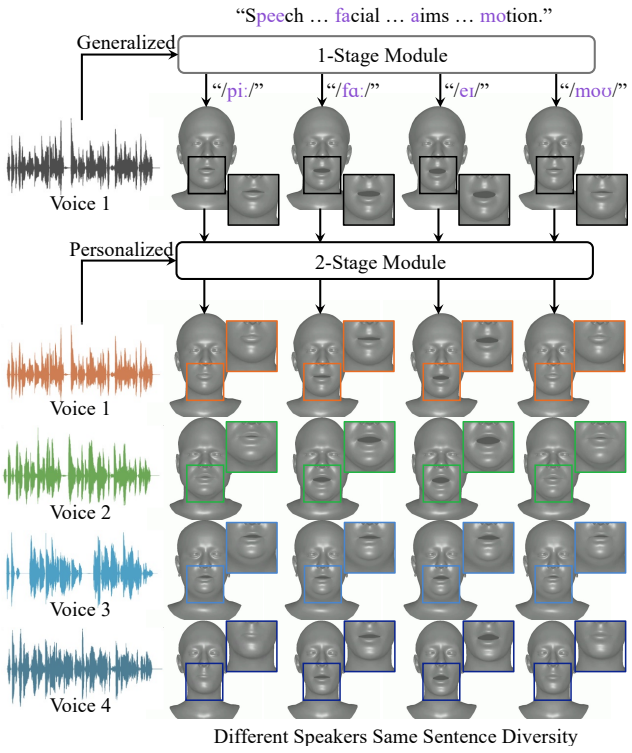


Figure S4. General Motion synchronized with audio is obtained in 1-stage (i.e. w/o 2-stage). Then, Personalized Motion is synthesized in 2-stage using speaking style features extracted from different speakers (i.e. w/ 2-stage).

1280-dimensional feature. After that, it goes through conv1d, group normalization, and ReLU in order, and then learns the features that appear throughout the frame through an average pooling layer. Finally, it is projected as a feature with a dimension of d_{txt} through a linear layer. [8] The Motion Memory uses a text addressing vector with a dimension of d_{txt} as a query and emits a motion feature with a dimension of d_m. The ASR encoder uses the HuBERT [18] structure and learned parameters. First, the audio feature is encoded with a 1D convolution layer and a group normalization layer. The GeLU nonlinear function is used during encoding. After that, linear interpolation is used to match the audio fps with the motion fps (30 fps for VOCASET and 25 fps for BIWI). After that, it passes through a transformer encoder, and then a pre-trained linear layer is used to map the feature to the vocab size. The Motion Decoder reduces the dimension to d_m using a linear layer, combining the retrieved motion feature and the text representation feature, and performs positional encoding. After that, the transformer decoder is used to induce the motion feature that matches the text representation. The obtained motion feature passes through a linear layer and synthesizes a motion that moves a neutral face mesh.

Competitors	Lip Sync(%)	Realism(%)
vs. GT	41.1	40.3

Table S6. User study: our method vs. GT on VOCASET [6].

E.2. Detail of User Study

In Figure S8, we attach the user study we provided to the subjects. We evaluated our method against other methods using the user study form used in the faceformer [12] and the mimic [19]. Our qualitative evaluation questionnaire consists of a total of 90 questions. Five questions are to check whether the participant is participating sincerely through the qualification video and to remove outliers. For the remaining 85 questions, we evaluate three qualitative metrics in total to compare the output of our *MemoryTalker* with the outputs of existing SoTA methods and GT for sentences randomly sampled from the 40 sentences.

For the evaluation of Realism, we induce the participant to choose A/B pairs by asking the following question: "Comparing the two full faces, which one looks more realistic?". In this case, participants see the full face in two samples and choose the more natural option.

To measure the Lip-sync, we conducted an A/B test with the question: "Comparing the two lips, which one looks more realistic?". As with evaluating realism, we asked participants to choose the sample that was more synchronized with the audio among the A/B samples. This allowed us to subjectively evaluate the degree of synchronization between the audio and the lip region. To measure how well our *MemoryTalker* captures the speaking style of the ground truth (GT), we compare whether our output is more similar to the GT than the outputs from other models. To this end, we randomly placed the GT video in the first position, and the videos to be compared in the second and third positions. And then, we asked the participant to answer the question "Comparing the speaking style (including the amplitude of mouth opening and closing, the dimensionality of pouting, etc.) of the last two faces, which one is more consistent with the first video?".

Additionally, to assess participants' reliability in the qualification question, we randomly place the same first-position video in either the second or the third position. At this time, the participant is asked a question about speaking style ("Comparing the speaking style (including the amplitude of mouth opening and closing, the dimensionality of pouting, etc.) of the last two faces, which one is more consistent with the first video?"), and this choice is different from the previous one in that it has a fixed correct answer. Therefore, if the participant answers this question incorrectly, we consider the participant an outlier and remove it from the statistics, thereby improving the quality of our evaluation method.

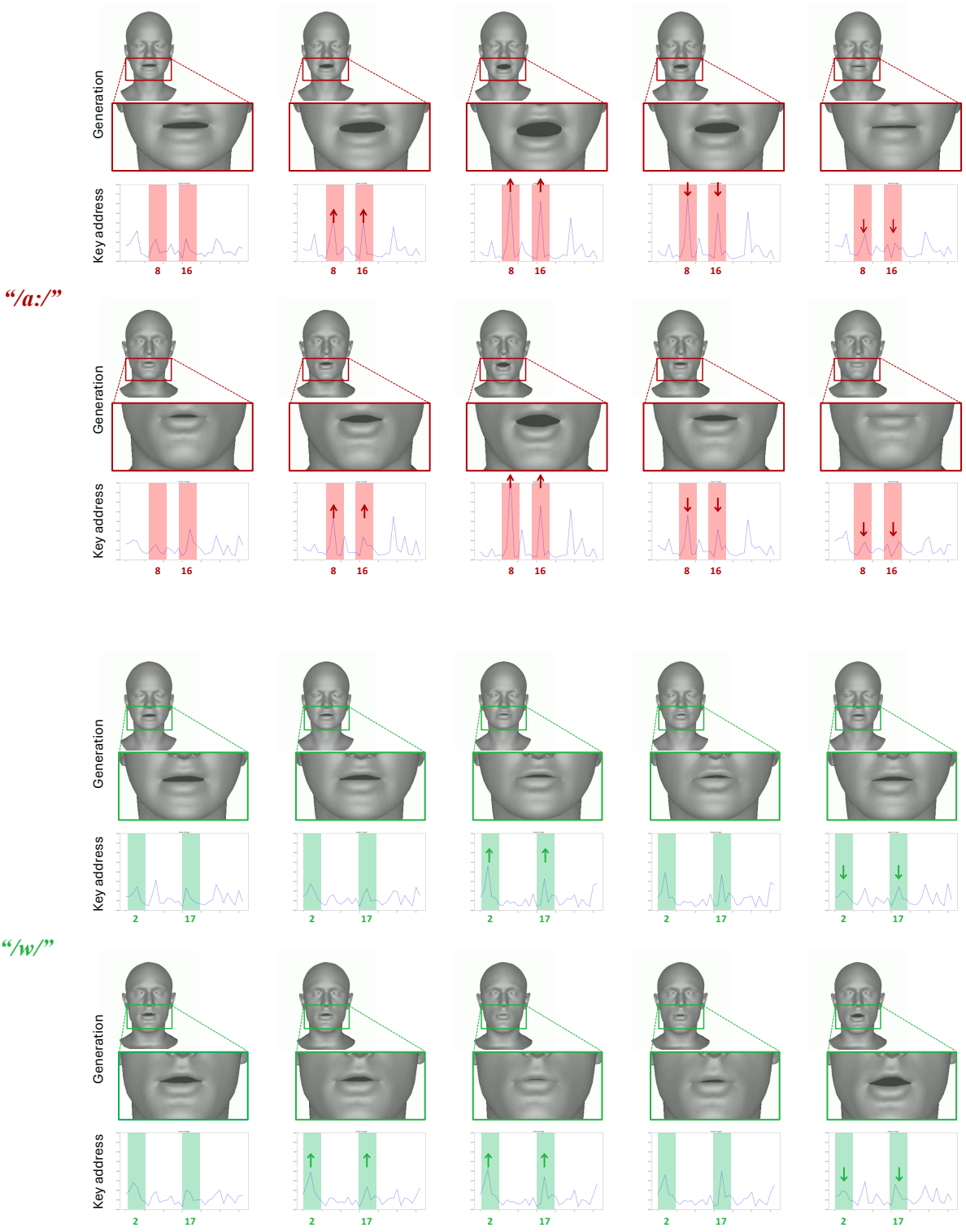


Figure S5. Key addresses from audio input and corresponding generated mesh in a sequence. Note that the key addressing vector here is from the ASR model, which is to extract speaker-neutral general motions. Our personalization is further applied to the memory components afterwards.

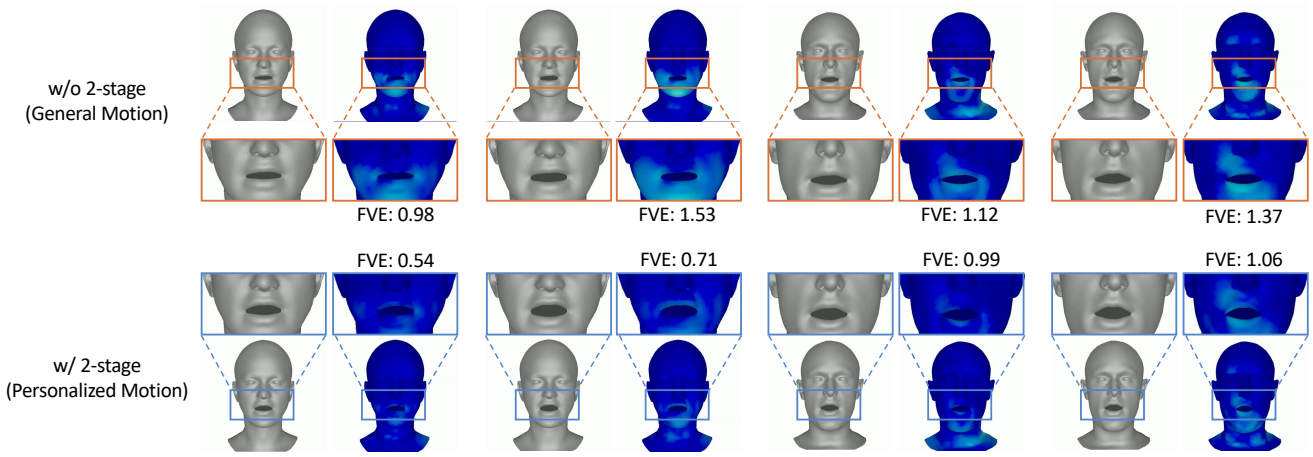


Figure S6. Qualitative comparison of results of learning with 1-stage only (general motion) and those of learning with 2-stage (personalized motion).

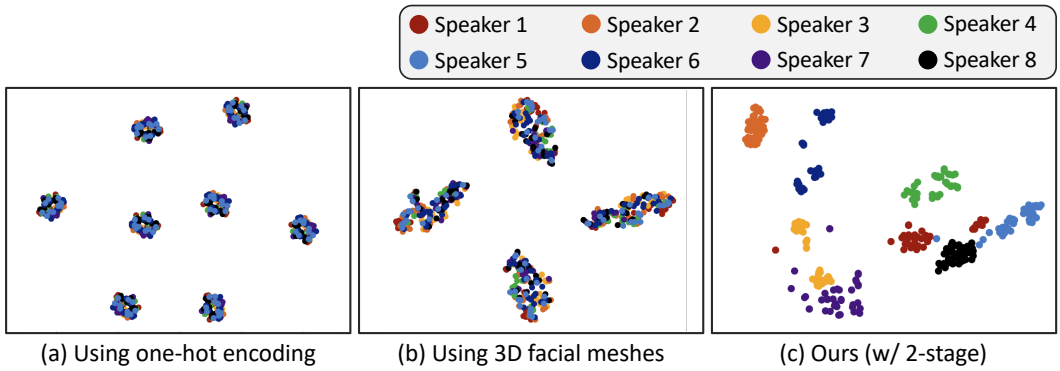
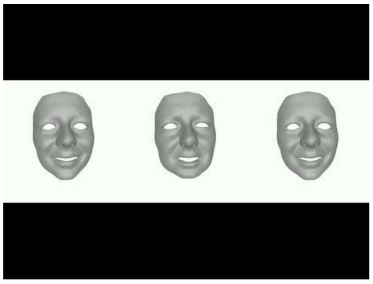


Figure S7. Feature visualization according to different style encoding types: one-hot, 3D facial meshes, and ours.

In summary, we evaluated three qualitative metrics (Speaking Style, Realism, and Lip-sync) by randomly selecting ten sentences per model, following a similar scale to previous study[12]. Based on the ten sampled videos for each model, we split them into two groups: five videos were used to evaluate Speaking Style, and the remaining five were used to evaluate Realism and Lip-sync. Specifically, we generated one question per video for Speaking Style (5 questions), and two questions per video (one for Realism, one for Lip-sync) for the other five videos (10 questions). Consequently, each model yielded a total of 15 questions (5 + 10) from its ten samples. With five models, we obtained 50 samples in total and generated 90 questions. Additionally, we selected ten more samples to directly compare our method with the ground truth (GT) as S6. Among these, five were used for qualification questions, and the remaining five were used for Realism and Lip-sync evaluations between our approach and the GT. All questions were administered in a forced-choice format, and any participant who answered even one qualification question incorrectly was excluded from the final statistics. Furthermore, we provided two reminders to participants: (1) to ensure their computer's sound was on while watching the videos, and (2) to note that one or two of the videos might be used for qualification purposes, such that random guessing could result in disqualification. Because each video is relatively short (4–7 seconds), we allowed participants to watch them repeatedly up to five times and even replay them additionally, ensuring they could carefully assess each sample. A total of 33 people participated in our user study, and 2 of them were removed because they did not pass the qualification questions we provided. In our experience, the participants took about 40 minutes to complete our user study.



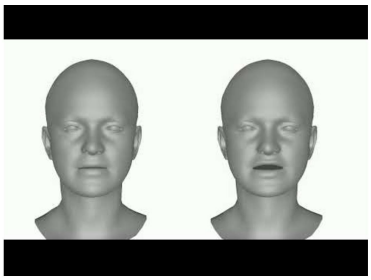
Q18. Please watch the video and answer the question. *

Comparing the **speaking style** (including the amplitude of mouth opening and closing, the dimensionality of pouting, etc.) of the last two faces, which one is more consistent with the first video?

Note: The first one on the left is a reference video.

- ○ The second
- ○ The third

(a) Qualified Questions



Q7. Please answer the following questions, after you watch the video. *

Comparing the two **full faces**, which one looks more realistic?

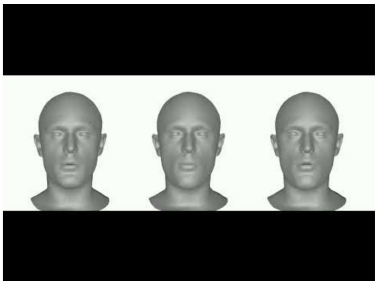
- ○ The Left one looks more realistic
- ○ The Right one looks more realistic

Q7. Please answer the following questions, after you watch the video. *

Comparing the two **lips**, which one looks more realistic?

- ○ The Left one is more in sync with audio
- ○ The Right one is more in sync with audio

(b) A/B Test (Realism / Lip-Sync)



Q8. Please watch the video and answer the question. *

Comparing the **speaking style** (including the amplitude of mouth opening and closing, the dimensionality of pouting, etc.) of the last two faces, which one is more consistent with the first video?

Note: The first one on the left is a reference video.

- ○ The second
- ○ The third

(c) A/B Test (Speaking Style)

Figure S8. Examples of conducted user study.