DiffusionTalker: Efficient and Compact Speech-Driven 3D Talking Head via Personalizer-Guided Distillation

Anonymous ICME submission

***Abstract*—Real-time speech-driven 3D facial animation has been attractive in academia and industry. Traditional methods mainly focus on learning a deterministic mapping from speech to animation. Recent approaches start to consider the nondeterministic fact of speech-driven 3D face animation and employ the diffusion model for the task. Existing diffusion-based methods can improve the diversity of facial animation. However, personalized speaking styles conveying accurate lip language is still lacking, besides, efficiency and compactness still need to be improved. In this work, we propose DiffusionTalker to address the above limitations via personalizer-guided distillation. In terms of personalization, we introduce a contrastive personalizer that learns identity and emotion embeddings to capture speaking styles from audio. We further propose a personalizer enhancer during distillation to enhance the influence of embeddings on facial animation. For efficiency, we use iterative distillation to reduce the steps required for animation generation and achieve more than 8x speedup in inference. To achieve compactness, we distill the large teacher model into a smaller student model, reducing our model's storage by 86.4% while minimizing performance loss. After distillation, users can derive their identity and emotion embeddings from audio to quickly create personalized animations that reflect specific speaking styles. Extensive experiments are conducted to demonstrate that our method outperforms state-of-the-art methods. The code will be released.**

***Index Terms*—Diffusion, Distillation, Efficiency, Compactness, Personalization**

I. Introduction

Speech-driven 3D facial animation is pivotal in applications such as virtual reality [1], augmented reality [2], and computer games [3], [4]. This technology involves predicting facial parameters from audio sequences to drive a 3D facial model for animation. Real-time speech-driven 3D animation of special interests in communication anywhere, such as on mobile devices or VR glasses, which places high demands on the accuracy of lip movement, efficiency of inference speed and the compactness of model parameters.

With the advent of deep learning, data-driven techniques for speech-driven 3D animation have gained popularity [5], [6], offering better performance. VOCA [7] utilized CNNs for speech-driven 3D face generation with 12 FLAME-based facial models [8], but ignored emotional speech impacts on expressions. EmoTalk [9], using Transformers [10], introduced an emotion disentangling encoder, but was limited by its reliance on user-defined one-hot identity encoding, restricting personalized speaking style capture. SelfTalk [11] uses a self-supervised approach, focusing on facial action units for better lip movement accuracy, but adds redundant encoders and decoders, increasing storage needs. Most methods employ deterministic networks, while speech-driven 3D facial animation is inherently nondeterministic [12]. FaceDiffuser [12] addresses this with a diffusion architecture, yielding impressive results, but its inference speed is slowed by a denoising process that requires thousands of steps.

Overall, current methods face three primary limitations. First, the slow inference speed of existing approaches is a significant challenge, particularly for real-time applications, with long-step diffusion-based methods being especially impacted. Second, many methods rely on large pre-trained audio encoders, such as HuBERT [13] and Wav2Vec [14], which introduces redundancy in model parameters and leads to excessive storage requirements. Third, most methods struggle to effectively integrate identity and emotion in 3D face animation. In practice, even when different people say the same words with the same emotions, their facial movements remain unique. And a person's expressions can vary greatly with their emotional state, even when repeating the same phrases.

To address existing limitations, we propose DiffusionTalker, an efficient and compact model that combines identity and emotion information to generate personalized 3D facial animations from speech, as illustrated in Fig. 1. A contrastive personalizer module is proposed, in which identity and emotion are extracted from the audio input through contrastive learning. Using a cross-attention mechanism, these embeddings are fused into a personalized embedding, which serves as a key conditioning factor for the denoising process. This personalized embedding guides the motion decoder in generating customized facial animations. To boost efficiency, a personalizer-guided distillation approach is used to iteratively distill a teacher model with N steps into a student model with n steps ($N > n$), thereby accelerating the inference speed. For model compression, we distill the large pre-trained audio encoder from the teacher model into a smaller encoder in the student model, significantly reducing both model parameters and storage requirements.

Our main contributions are summarized as follows.

- We introduce DiffusionTalker, an efficient and compact 3D face diffuser that generates personalized 3D facial animations based on the diffusion model.
- A personalizer-guided distillation method is employed to iteratively reduce the denoising steps, significantly accelerating inference and achieving a speed-up of more than 8x. Furthermore, this method distills a large model

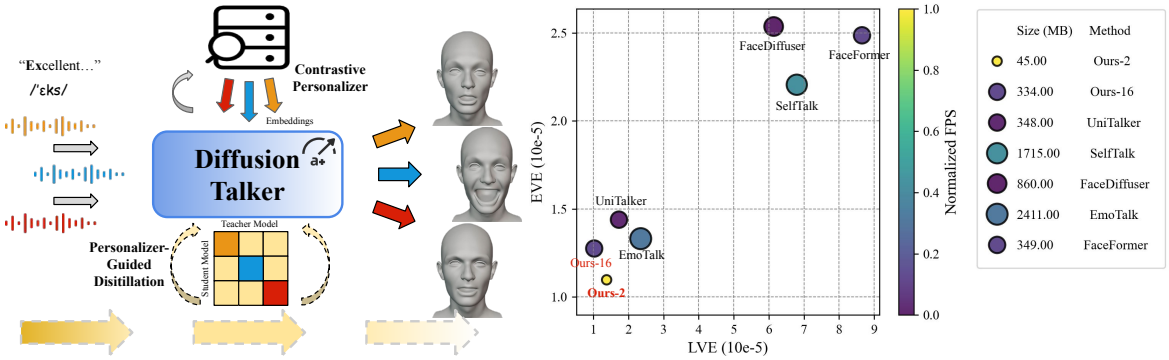

Fig. 1. **The illustration of DiffusionTalker.** We reduce the steps of the diffusion model for faster inference and compress the model size for compactness by personalizer-guided distillation. Our distilled 2-step model surpasses the state-of-the-art methods in terms of emotional expression and lip accuracy, while also achieving the fastest inference speed and the fewest model parameters.

into a smaller one, resulting in 86.4% reduction in model size.

- We introduce a contrastive personalizer that learns identity and emotion embeddings using contrastive learning, which are then integrated into a personalized embedding through a cross-attention mechanism.
- A personalizer enhancer is applied during distillation to bring personalized embeddings of the same identity-emotion closer together and to push those of different identity-emotions farther apart, further enhancing the model's ability to generate personalized facial animations.

II. Methodology

A. Preliminaries

The vanilla Denoising Diffusion Probabilistic Model (DDPM) [15] consists of two processes: the diffusion process and the denoising process. The diffusion process gradually adds Gaussian noise to clean data $\boldsymbol{x}_0$ using a Markov chain, resulting in a noisy distribution $q(\boldsymbol{x}_T|\boldsymbol{x}_0)$:

$$q(\boldsymbol{x}_T|\boldsymbol{x}_0) = \prod_{t=1}^{T} q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) \tag{1}$$

where T is the diffusion step. Specifically, the clean data $\boldsymbol{x}_0$ is noised to $\boldsymbol{x}_t$, as shown in the following equation:

$$\boldsymbol{x}_t = \sqrt{\overline{\alpha_t}}\boldsymbol{x}_0 + \sqrt{1-\overline{\alpha_t}}\boldsymbol{\epsilon} \tag{2}$$

where $\overline{\alpha_t}$ is hyperparameters, and $\boldsymbol{\epsilon}$ is Gaussian noise.

Denoising is an inverse stepwise process $q'(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)$, transitioning from $\boldsymbol{x}_t$ to $\boldsymbol{x}_{t-1}$, as shown in the following equation:

$$\boldsymbol{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(\boldsymbol{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t)\right) + \sigma_t \mathbf{z} \tag{3}$$

where $\boldsymbol{\epsilon}_\theta$ is the network to predict noise, θ is the parameters, and $\sigma_t\mathbf{z}$ is a noise term used to introduce diversity. Similarly, the denoising process also involves T progressive steps.

B. Overview

DiffusionTalker uses a DDPM model, conditioning on audio, identity, and emotion to guide the denoising process for generating speech-driven 3D facial animations. As shown in Fig. 2, the model consists of the contrastive personalizer and the motion decoder. The personalizer extracts audio features, identity, and emotion embeddings from the input speech, combining them into a personalized embedding. The motion decoder uses these embeddings, along with the noisy facial animation $\boldsymbol{x}_t$ and time step t, to iteratively remove noise and predict the final facial animation $\hat{\boldsymbol{x}}$. The model is trained by minimizing the reconstruction loss $\mathcal{L}_{rec}$ between the predicted animation and the ground truth, as shown in the following loss function:

$$\mathcal{L}_{rec} = \|\boldsymbol{x}_0 - \mathrm{DT}_\theta(\boldsymbol{a}, \boldsymbol{i}, \boldsymbol{e}, \boldsymbol{x}_t, t)\|^2 \tag{4}$$

where DT is our model, θ represents the model parameters, $\boldsymbol{a}$ is the audio sequence, $\boldsymbol{i}$ and $\boldsymbol{e}$ are the embeddings of identity and emotion, respectively, and $\boldsymbol{x}_0$ is the ground truth.

For personalizer-guided distillation, we use an N-step model as the teacher model and an n-step (where $N = 2n$) model as the student. This reduces inference time and compresses the audio encoder to optimize storage and performance, while maintaining lip accuracy. The personalizer enhancer further strengthens the representational power of the embeddings through contrastive learning.

C. Contrastive Personalizer

To equip the model with personalization ability, we propose a contrastive personalizer to extract and integrate identity and emotion from the input speech using contrastive learning similar to [16]. As shown in Fig. 3, we establish an identity embedding library and an emotion library. Each speaker is associated with an identity one, $\boldsymbol{i}$, and each type of emotion corresponds to an emotion embedding, $\boldsymbol{e}$.

The aim of this module is to formulate the relationship between identity/emotion embeddings and audio sequences. The audio sequence is first fed into the audio encoder for its feature, $\boldsymbol{F}_a$. Meanwhile, the libraries of identity and emotion are processed separately by the encoders of both to generate the corresponding features $\boldsymbol{F}_{id}$ and $\boldsymbol{F}_e$. These features are then used to calculate the contrastive loss with F_a separately, which is adopted to train the parameters in this module.

Specifically, we first apply L2 normalization to both the audio feature and identity feature. Next, we perform matrix multiplication on these normalized features and the results are used to compute the cross-entropy loss with one-hot labels. The labels are constructed based on the positive or negative

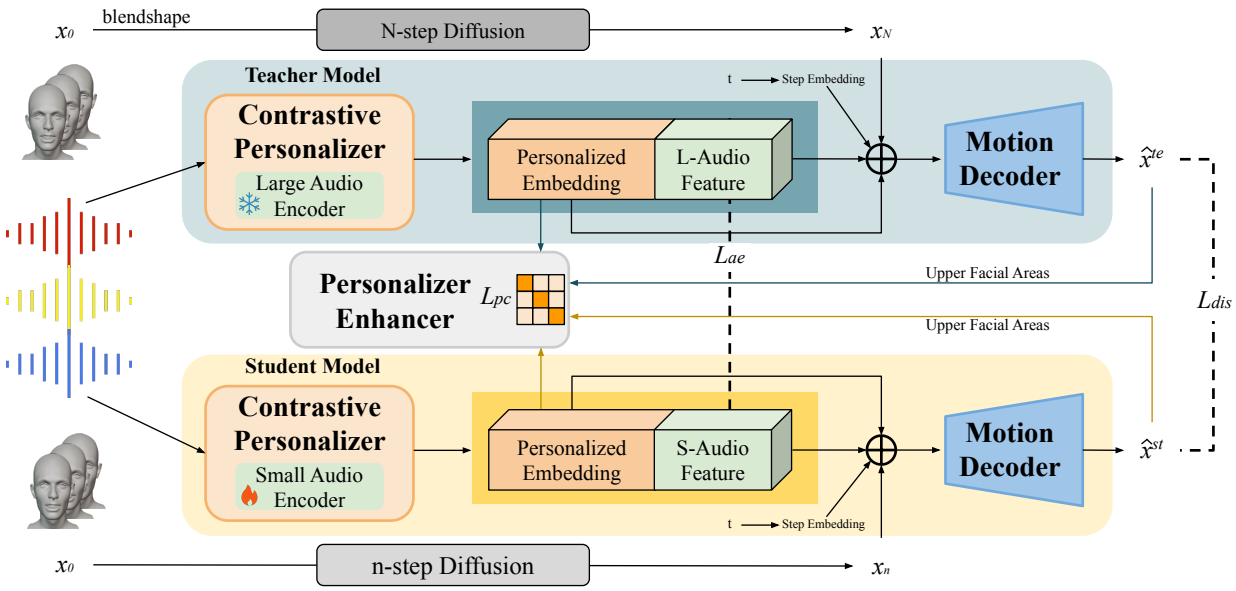

Fig. 2. **Pipeline of DiffusionTalker.** DiffusionTalker employs a contrastive personalizer to extract audio features and personalized embeddings from the input speech. These representations serve as conditioning inputs to guide the motion decoder in denoising noisy facial animations effectively. In personalizer-guided distillation process, the number of steps in the student model is iteratively reduced to half of the original, significantly accelerating inference. Simultaneously, the model parameters are compressed to create a more compact and efficient model. The personalizer enhancer integrates the personalized embedding with the facial areas of the predicted results, leveraging contrastive learning to strengthen the embedding's representational capacity.

sample relationship between the audio feature and identity feature. For identity, $\boldsymbol{F}_a$ has only one positive sample, which is $\boldsymbol{F}_{id}$ that corresponds to the current identity label. All other embeddings are regarded as negative samples. The contrastive loss between $\boldsymbol{F}_{id}$ and $\boldsymbol{F}_a$ is formulated as:

$$\mathcal{L}_{con}^{\boldsymbol{F}_{id}-\boldsymbol{F}_a} = -\log \frac{exp(\boldsymbol{F}_a^\top \boldsymbol{F}_{id}^+/\tau)}{\sum_{k=1}^{M_{id}} exp(\boldsymbol{F}_a^\top \boldsymbol{F}_{id}^k/\tau)} \tag{5}$$

where M_{id} is the number of identity features, $\boldsymbol{F}^+$ is the positive sample of $\boldsymbol{F}_a$, and τ is a temperature hyperparameter.

For emotion, the formula of contrastive loss $L_{con}^{\boldsymbol{F}_e-\boldsymbol{F}_a}$ between $\boldsymbol{F}_e$ and $\boldsymbol{F}_a$ is similar to $L_{con}^{\boldsymbol{F}_{id}-\boldsymbol{F}_a}$, which is shown as:

$$\mathcal{L}_{con}^{\boldsymbol{F}_e-\boldsymbol{F}_a} = -\log \frac{exp(\boldsymbol{F}_a^\top \boldsymbol{F}_e^+/\tau)}{\sum_{k=1}^{M_e} exp(\boldsymbol{F}_a^\top \boldsymbol{F}_e^k/\tau)} \tag{6}$$

where M_e is the number of emotion features.

Simultaneously, the contrastive personalizer searches the libraries to retrieve $\boldsymbol{i}$ and $\boldsymbol{e}$ that match the current audio according to the input emotion and identity labels. $\boldsymbol{i}$ and $\boldsymbol{e}$ are then processed by the personalized integrator, which employs a cross-attention mechanism to merge them into a singular personalized embedding, $\boldsymbol{p}$. This composite embedding is then utilized as a key input for the decoder. The personalized integrator is illustrated as follows:

$$\boldsymbol{p} = \text{softmax}\left(\frac{\boldsymbol{i} \cdot f_k(\boldsymbol{e})^T}{\sqrt{d_k}}\right) \cdot f_v(\boldsymbol{e}) \tag{7}$$

where $\boldsymbol{i}$ is identity embedding, $\boldsymbol{e}$ is emotion embedding, f_k and f_v are MLPs, and d_k is key dimension.

During inference, the personalizer can autonomously infer the emotion and identity embeddings that are most similar to the input audio sequence, without the need for labels.

The motion decoder processes a concatenated input consisting of the personalized embedding $\boldsymbol{p}$, audio feature $\boldsymbol{F}_a$, noisy

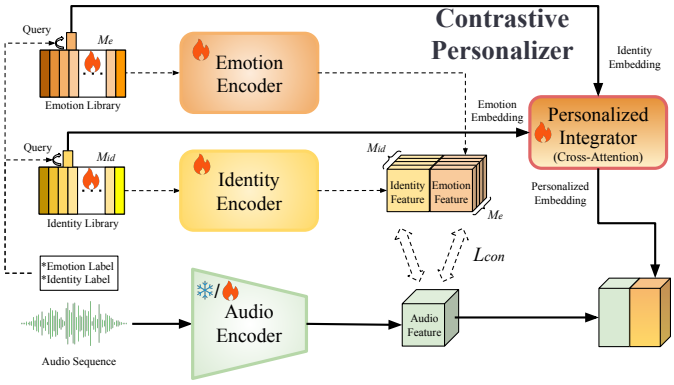

Fig. 3. **Contrastive Personalizer.** Contrastive learning is performed separately on audio features and identity/emotion features, and specific identity/emotion embeddings are fused into personalized embeddings.

animation $\boldsymbol{x}_t$, and time step t. It performs denoising from step t to step $t-1$ to generate the predicted facial animation, represented as $q'_m(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{F}_a, \boldsymbol{p}, t)$.

To capture the temporal dynamics of speech, we use a GRU backbone, which is compact and efficient. The input is first normalized via layer normalization, then processed through three GRU layers to capture contextual information, and finally passed through an MLP to generate the facial animation.

D. Personalizer-Guided Distillation

1) Acceleration: To accelerate inference and ensure the high quality of the generated animations, we utilize this distillation manner for fewer denoising steps. Inspired by [17], we introduce a student model $\hat{s}_\eta$ with n steps to match the pre-trained teacher model $\hat{t}_\theta$ with N steps (where $N = 2n$). We use $\boldsymbol{c}_t = (\boldsymbol{p}, \boldsymbol{F}_a, t)$ as the input of condition at time step t. Specifically, we perform one DDPM step for the student n to match two DDPM steps for the teacher $2n$ and $2n-1$.

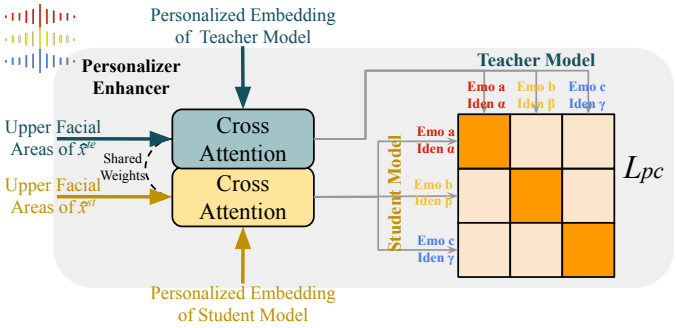

Fig. 4. **Personalizer Enhancer.** It is used to enhance the personalization.

Assume that the signal-to-noise ratio at time step $\tau \sim \mathcal{U}(0, n)$ is $\alpha_\tau^2/\sigma_\tau^2$, then the corresponding data $\boldsymbol{x}_0$ added with noise $\boldsymbol{\epsilon}$ is $\boldsymbol{x}_\tau = \alpha_\tau \boldsymbol{x}_0 + \sigma_\tau \boldsymbol{\epsilon}, \boldsymbol{\epsilon} \sim N(0, I)$.

For the teacher model, we compute the denoised output of the $2n$ and $2n-1$ steps. Thus, we can obtain noisy data $\boldsymbol{x}_{\tau'}$ at time step $\tau' = 2\tau$ as:

$$\hat{\boldsymbol{x}}_{\tau'} = \alpha_{\tau'} \hat{t}_\theta(\boldsymbol{x}_\tau, \boldsymbol{c}_{\tau'}) + \frac{\sigma_{\tau'}}{\sigma_\tau}(\boldsymbol{x}_\tau - \alpha_\tau \hat{t}_\theta(\boldsymbol{x}_\tau, \boldsymbol{c}_{\tau'})) \quad (8)$$

At step $\tau'' = 2\tau - 1$, noisy data $\boldsymbol{x}_{\tau''}$ can be calculated as:

$$\hat{\boldsymbol{x}}_{\tau''} = \alpha_{\tau''} \hat{t}_\theta(\hat{\boldsymbol{x}}_{\tau'}, \boldsymbol{c}_{\tau''}) + \frac{\sigma_{\tau''}}{\sigma_{\tau'}}(\hat{\boldsymbol{x}}_{\tau'} - \alpha_{\tau'} \hat{t}_\theta(\hat{\boldsymbol{x}}_{\tau'}, \boldsymbol{c}_{\tau''})) \quad (9)$$

The final target $\hat{\boldsymbol{x}}^{te}$ from teacher model can be formulated as:

$$\hat{\boldsymbol{x}}^{te} = \frac{\hat{\boldsymbol{x}}_{\tau''} - (\sigma_{\tau''}/\sigma_\tau)\boldsymbol{x}_\tau}{\alpha_{\tau''} - (\sigma_{\tau''}/\sigma_\tau)\alpha_\tau} \quad (10)$$

For the student model, we initialize it with the same parameters as the teacher. We then add noise according to the student DDPM for τ steps. This noisy data is then fed into the student, resulting in the output denoted as $\hat{\boldsymbol{x}}^{st} = \hat{s}_\eta(\boldsymbol{x}_\tau, \boldsymbol{c}_\tau)$.

We apply $\hat{\boldsymbol{x}}^{te}$ to supervise the student model. The distill loss in this process is as follows:

$$\mathcal{L}_{dis} = max(\frac{\alpha_\tau^2}{\sigma_\tau^2}, 1)\|\hat{\boldsymbol{x}}^{te} - \hat{\boldsymbol{x}}^{st}\|^2 \quad (11)$$

Finally, DiffusionTalker can be distilled to half the original number of steps while preserving high generation quality.

2) Compression: During distillation, we assign a large pre-trained audio encoder to the teacher model to fully extract audio features, resulting in the large audio feature $\boldsymbol{F}_a^L$. For the student model, we assign a small audio encoder to obtain the small audio feature $\boldsymbol{F}_a^S$. We use an audio encoder (AE) loss based on cosine similarity to optimize the small audio encoder, aligning $\boldsymbol{F}_a^S$ with $\boldsymbol{F}_a^L$. The formula is as follows:

$$\mathcal{L}_{\text{ae}} = 1 - \text{mean}\left(\frac{\boldsymbol{F}_a^L}{\|\boldsymbol{F}_a^L\|_2} \cdot \frac{\boldsymbol{F}_a^S}{\|\boldsymbol{F}_a^S\|_2}\right) \quad (12)$$

The large audio encoder employs the pre-trained HuBERT model [13], while the small one utilizes the feature extractor (CNN) from HuBERT to process audio input. Subsequently, it extracts temporal features through a bidirectional LSTM and maps them to the target dimension using a fully connected layer. Finally, the distilled student model significantly reduces model size while maintaining lip accuracy.

3) Personalizer Enhancer: To further enhance the influence of identity and emotion on the mapping from speech to facial animation, we propose a personalizer enhancer during the distillation process.

As illustrated in Fig. 4, we set the batch size as $M = 3$. Specifically, the red speech icon represents an audio sequence corresponding to identity a with emotion α, the yellow icon denotes identity b with emotion β, and the blue icon represents identity c with emotion γ, where $\alpha \neq \beta \neq \gamma, a \neq b \neq c$. In the context of the teacher model, the audio sequences are passed through the personalizer that generates three distinct personalized embeddings. Each personalized embedding $\boldsymbol{p}$ is then employed in cross-attention with the upper facial areas of the predicted results $\hat{\boldsymbol{x}}_u^{te}$, resulting in three personalized feature maps $\boldsymbol{F}_{pu}^{te}$, for the teacher model, where $f(\cdot)$ is MLP.

$$\boldsymbol{F}_{pu}^{te} = \text{softmax}\left(\frac{\boldsymbol{p} \cdot f_k(\hat{\boldsymbol{x}}_u^{te})^T}{\sqrt{d_k}}\right) \cdot f_v(\hat{\boldsymbol{x}}_u^{te}) \quad (13)$$

A similar procedure is applied to the student model, yielding three corresponding feature maps $\boldsymbol{F}_{pu}^{st}$. The personalized feature maps $\boldsymbol{F}_{pu}^{te}$ and $\boldsymbol{F}_{pu}^{st}$ are then compared using a contrastive learning method, which computes the personalized contrastive (PC) loss $\mathcal{L}_{pc}$. The formula is shown as:

$$\mathcal{L}_{pc} = -\frac{1}{M}\sum_{i=1}^{M} \log \frac{exp(\boldsymbol{F}_{pu}^{te\ i\ \top} \boldsymbol{F}_{pu}^{st\ +}/\tau)}{\sum_{j=1}^{M} exp(\boldsymbol{F}_{pu}^{te\ i\ \top} \boldsymbol{F}_{pu}^{st\ j}/\tau)} \quad (14)$$

Notably, the teacher and student models share identical weight parameters for the cross-attention mechanism. More details are provided in the supplementary materials.

This approach effectively amplifies the contribution of personalized embeddings to the facial animation output, thereby improving the model's capacity to capture and reflect identity and emotion in the generated animation.

E. Training

The training process begins by using $\mathcal{L}_{tea}$ to train an initial teacher model until convergence. The formula is shown as:

$$\mathcal{L}_{tea} = \lambda_1 \mathcal{L}_{rec} + \lambda_2 \mathcal{L}_{con}^{\boldsymbol{F}_i - \boldsymbol{F}_a} + \lambda_2 \mathcal{L}_{con}^{\boldsymbol{F}_e - \boldsymbol{F}_a} \quad (15)$$

This teacher model is then used for distillation to train a half-step student model using $\mathcal{L}_{stu}$. The student model is subsequently promoted to the role of the teacher, facilitating iterative distillation to train a new half-step student model with $\mathcal{L}_{stu}$, which is shown as:

$$\mathcal{L}_{stu} = \mathcal{L}_{tea} + \lambda_3 \mathcal{L}_{dis} + \lambda_4 \mathcal{L}_{ae} + \lambda_5 \mathcal{L}_{pc} \quad (16)$$

where λ_1 = 1.0, λ_2 = 0.007, λ_3 = 0.1, λ_4 = 1, λ_5 = 0.05 are fixed in all our experiments.

The trainable components of DiffusionTalker include the identity and emotion library, identity and emotion encoder, personalized integrator, small audio encoder, step embedding, motion decoder, and personalizer enhancer, while the parameters of the large audio encoder remain fixed. During distillation, the teacher's parameters are fixed, while the student's parameters are trainable.

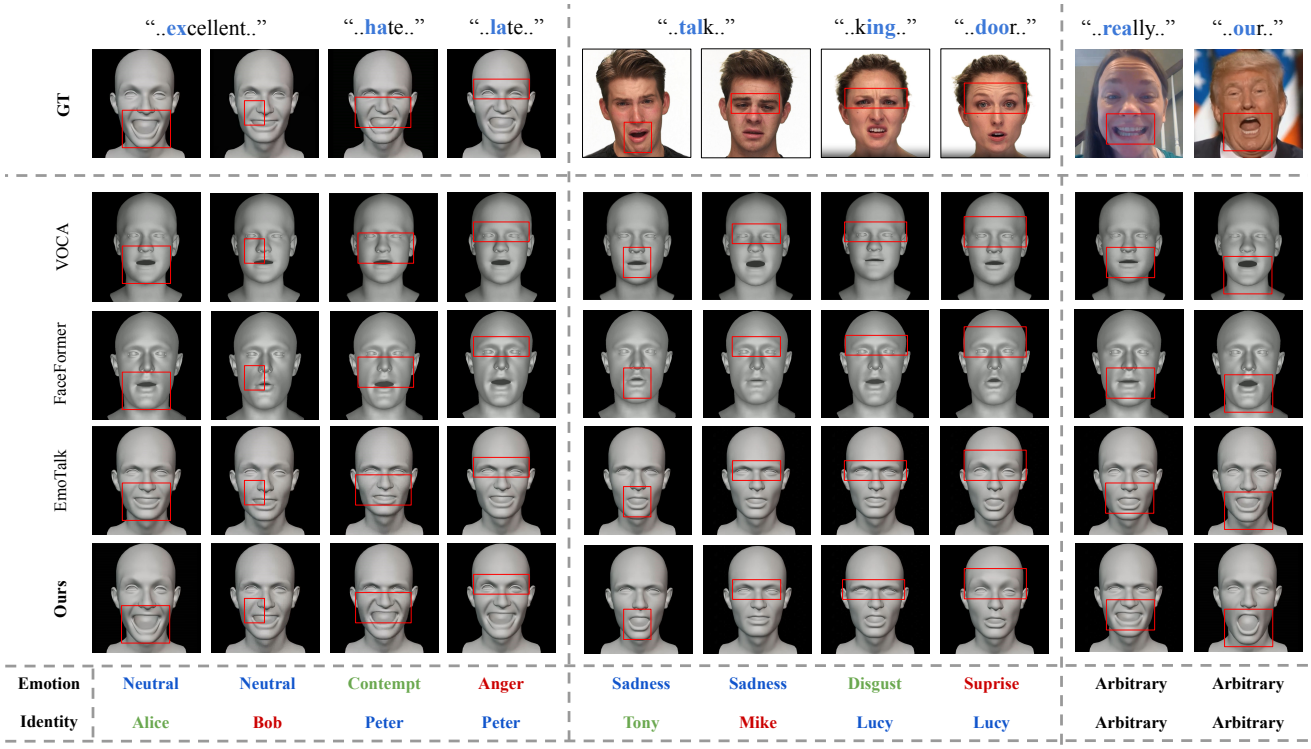

Fig. 5. Qualitative comparisons with other methods on BEAT-Test(left), 3D-ETF(middle), and in-the-wild videos(right). We input speech with different identities and emotions into various models and present the same frames to compare them with the ground truth (GT). As indicated by the red box, it can be witnessed that on the first two datasets, our model accurately discerns facial action changes among various identities and precisely generates facial expressions corresponding to specific emotions. Even on in-the-wild videos, our model can produce accurate results.

III. Experiments

A. Datasets and Metrics

To evaluate the robustness of DiffusionTalker, we select three datasets: BEAT [18], 3D-ETF(RAVDESS) [9], and VOCASET [7]. Our method and the baselines are trained on the first two datasets. DiffusionTalker conducts zero-shot testing on the third dataset. We follow the same data preprocessing as the baselines and use the EmoTalk's bs2FLAME [9] method to align the data. Please refer to the supplementary materials for more details on these datasets.

To evaluate the performance of DiffusionTalker, we use five metrics: LVE, EVE, FDD, FPS, and model size. **LVE**: Lip vertex error, measuring the maximum l_2 error of the predicted vs. ground-truth lip areas. **EVE**: Emotional vertex error, assessing the maximum error l_2 for the coefficients of the eyes, eyebrows, forehead, and surrounding areas. **FDD**: Facial dynamics deviation [19], quantifying upper-face dynamic differences; Lower values indicate more natural animations. **FPS**: Frames Per Second, indicating inference speed; higher values mean faster frame generation. **Model size** represents the number of parameters in the model.

B. Experimental Results

a) Animation synthesis comparison.: We test our DiffusionTalker model against other advanced methods on the 3D-ETF dataset with EVE, LVE, FDD, FPS and model size metrics, as shown in Tab. I. Our 16-step method (ours-16) serves as the initial teacher model and achieves the best performance across the LVE and FDD metrics, while our distilled 2-step model (ours-2) incurs minimal performance loss, achieving the second-best LVE and FDD results. Due to the effect of the personalizer enhancer, the EVE metric for emotion evaluation has decreased, indicating that the model's ability to express personalization has become stronger. Moreover, ours-2 achieves more than 8x speed-up in inference and an 86.4% reduction in model size. Compared to baselines, our 2-step model achieves the fastest speed and the smallest model size while delivering the best accuracy in lip region prediction (LVE). In terms of the naturalness of facial dynamic changes (FDD), our method also outperforms existing approaches. Our method shows a significant advantage on the EVE metric, indicating that the personalized embeddings learned through the contrastive personalizer and personalizer-guided distillation make a substantial contribution to emotional representation.

TABLE I
THE QUANTITATIVE COMPARISON ON THE 3D-ETF. BEST RESULTS IN BOLD, SECOND-BEST UNDERLINED.

Dataset	Method	EVE ↓ ($\times 10^{-5}$)	LVE ↓ ($\times 10^{-5}$)	FDD ↓ ($\times 10^{-7}$)	FPS↑	Size(MB)↓
3D-ETF (Test)	FaceFormer	2.487	8.656	11.909	426.01	349
	EmoTalk	1.331	2.347	2.196	1063.13	2411
	FaceDiffuser	2.537	6.137	8.172	10.33	860
	SelfTalk	2.206	6.796	11.622	<u>1428.57</u>	1715
	UniTalker-B	1.439	1.727	3.004	68.75	348
	Ours-16	<u>1.275</u>	**1.021**	**2.023**	440.01	<u>334</u>
	Ours-2	**1.097**	<u>1.375</u>	<u>2.177</u>	**3632.15**	**45**

b) Zero-shot evaluation.: To test generalizability across all methods, we test models on VOCASET and calculate the LVE and FDD, as shown in Tab. II. Our approach demonstrates

TABLE II
THE ZERO-SHOT TEST ON THE VOCASET.

Dataset	Method	LVE ↓ ($\times 10^{-5}$)	FDD ↓ ($\times 10^{-7}$)	Zero Shot
VOCASET (Test)	FaceFormer	1.170	2.493	✗
	FaceDiffuser	0.973	1.754	✗
	SelfTalk	0.967	**1.049**	✗
	UniTalker-B	**0.814**	1.396	✗
	Ours-2	<u>0.857</u>	<u>1.198</u>	✓

TABLE III
ABLATION STUDY FOR OUR 2-STEP MODEL ON 3D-ETF.

Settings	EVE↓ ($\times 10^{-5}$)	LVE ↓ ($\times 10^{-5}$)	FDD ↓ ($\times 10^{-7}$)
Ours-2	**1.097**	1.375	**2.177**
w/o identity embedding	1.223	1.412	2.301
w/o emotion embedding	1.968	1.547	2.570
w/o enhancer	1.712	**1.116**	2.512
w/o distillation	1.305	1.647	2.618

robust generalization capabilities in zero-shot scenarios and is highly competitive compared to state-of-the-art models.

c) Ablation study.: We conduct an ablation study to assess the impact of each component. In Tab. III, the emotion embedding and enhancer have the greatest impact on EVE, while the enhancer's focus on the upper face leads to a slight degradation in lip region accuracy. In general, our distillation method enables the student model to learn effective knowledge from the teacher model, significantly improving model performance.

d) Visualization comparison.: Visualizations generated by the model are crucial for performance evaluation. We test all methods on BEAT-Test, 3D-ETF, and in-the-wild videos to compare facial animations. By feeding speech with diverse identities and emotions into the models, we show side-by-side comparisons with the ground truth (GT) in Fig. 5. Our model demonstrates the highest sensitivity to varying identities. For example, in columns 1, 2, 5, and 6, when given audio from different speakers expressing the same emotion, it produces results closest to the GT, especially in the lip movements. When the same identity speaks with different emotions (columns 3, 4, 7, and 8), the model accurately reflects these emotional changes. Finally, when given speech with any identity and emotion, our model generates personalized facial animations, as seen in columns 9 and 10.

IV. CONCLUSION

In this work, we introduced DiffusionTalker, a compact and efficient diffusion model to generate personalized 3D facial animations from speech. By combining identity and emotion embeddings and utilizing a personalizer-guided distillation approach, we significantly improve inference speed and reduce model size. The results show that DiffusionTalker delivers superior performance, making it suitable for real-time applications in virtual and augmented reality. This work provides a step forward in creating personalized and efficient models for interactive 3D facial animation.

REFERENCES

[1] Isabell Wohlgenannt, Alexander Simons, and Stefan Stieglitz, "Virtual reality," *Business & Information Systems Engineering*, vol. 62, pp. 455–461, 2020.

[2] Gérard Chollet, Anna Esposito, Annie Gentes, Patrick Horain, Walid Karam, Zhenbo Li, Catherine Pelachaud, Patrick Perrot, Dijana Petrovska-Delacrétaz, Dianle Zhou, et al., "Multimodal human machine interactions in virtual and augmented reality," *Multimodal Signals: Cognitive and Algorithmic Issues: COST Action 2102 and euCognition International School Vietri sul Mare, Italy, April 21-26, 2008 Revised Selected and Invited Papers*, pp. 1–23, 2009.

[3] Heng Yu Ping, Lili Nurliyana Abdullah, Puteri Suhaiza Sulaiman, and Alfian Abdul Halin, "Computer facial animation: A review," *International Journal of Computer Theory and Engineering*, vol. 5, no. 4, pp. 658, 2013.

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

[5] 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*, 2022, pp. 18770–18780.

[6] Xiangyu Fan, Jiaqi Li, Zhiqian Lin, Weiye Xiao, and Lei Yang, "Unitalker: Scaling up audio-driven 3d facial animation through a unified model," *arXiv preprint arXiv:2408.00762*, 2024.

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

[8] 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.*, vol. 36, no. 6, pp. 194–1, 2017.

[9] 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*, 2023, pp. 20687–20697.

[10] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[11] 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*, 2023, pp. 5292–5301.

[12] Stefan Stan, Kazi Injamamul Haque, and Zerrin Yumak, "Facediffuser: Speech-driven 3d facial animation synthesis using diffusion," *arXiv preprint arXiv:2309.11306*, 2023.

[13] 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*, vol. 29, pp. 3451–3460, 2021.

[14] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12449–12460, 2020.

[15] Jonathan Ho, Ajay Jain, and Pieter Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.

[16] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al., "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.

[17] Tim Salimans and Jonathan Ho, "Progressive distillation for fast sampling of diffusion models," *arXiv preprint arXiv:2202.00512*, 2022.

[18] Haiyang Liu, Zihao Zhu, Naoya Iwamoto, Yichen Peng, Zhengqing Li, You Zhou, Elif Bozkurt, and Bo Zheng, "Beat: A large-scale semantic and emotional multi-modal dataset for conversational gestures synthesis," in *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VII*, Berlin, Heidelberg, 2022, p. 612–630, Springer-Verlag.

[19] 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*, 2023, pp. 12780–12790.