Model See Model Do: Speech-Driven Facial Animation with Style Control

YIFANG PAN, University of Toronto, Canada
KARAN SINGH, University of Toronto, Canada
LUIZ GUSTAVO HAFEMANN, Ubisoft La Forge, Canada

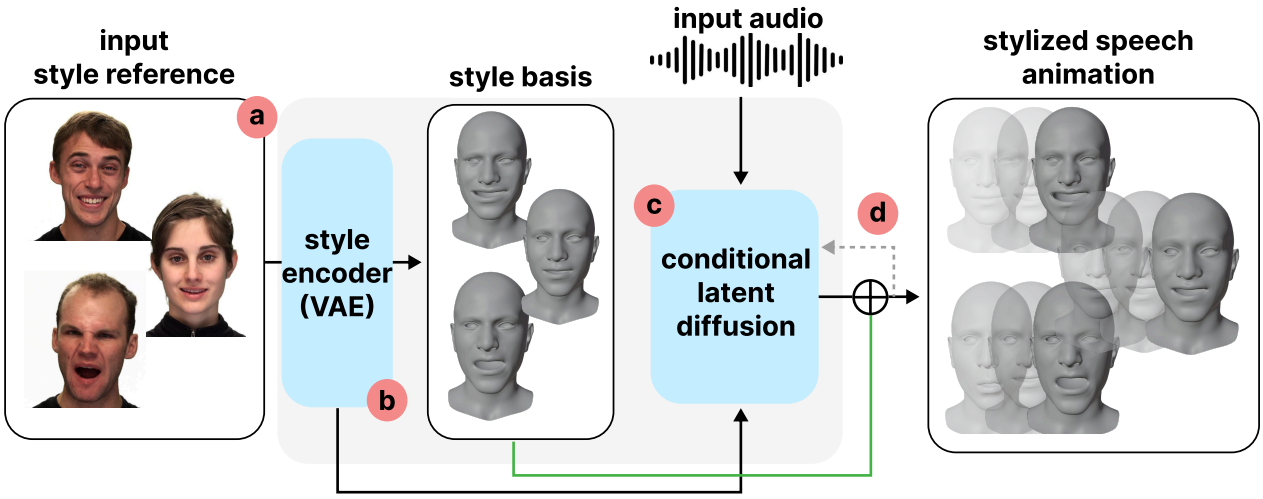

Fig. 1. We present an example-based system for generating stylistic 3D facial animations: (a) Given an arbitrary style reference, a style encoder (b) obtains latent style features and a style basis that reflects key poses from the reference. (c) A diffusion module, conditioned on audio and style features, produces the primary motion. (d) The style basis guide the primary motion throughout the diffusion process, gradually refining it at each diffusion step to produce the final animation.

Speech-driven 3D facial animation plays a key role in applications such as virtual avatars, gaming, and digital content creation. While existing methods have made significant progress in achieving accurate lip synchronization and generating basic emotional expressions, they often struggle to capture and effectively transfer nuanced performance styles. We propose a novel example-based generation framework that conditions a latent diffusion model on a reference style clip to produce highly expressive and temporally coherent facial animations. To address the challenge of accurately adhering to the style reference, we introduce a novel conditioning mechanism called style basis, which extracts key poses from the reference and additively guides the diffusion generation process to fit the style without compromising lip synchronization quality. This approach enables the model to capture subtle stylistic cues while ensuring that the generated animations align closely with the input speech. Extensive qualitative, quantitative, and perceptual evaluations demonstrate the effectiveness of our method in faithfully reproducing the desired style while achieving superior lip synchronization across various speech scenarios.

CCS Concepts: • **Facial Animation**; • **Animation System**;

Additional Key Words and Phrases: facial animation, diffusion model, animation system

ACM Reference Format:
Yifang Pan, Karan Singh, and Luiz Gustavo Hafemann. 2025. Model See Model Do: Speech-Driven Facial Animation with Style Control. In *Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers (SIGGRAPH Conference Papers '25), August 10–14, 2025, Vancouver, BC, Canada.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3721238.3730672

Authors' Contact Information: Yifang Pan, evan.pan@mail.utoronto.ca, University of Toronto, Toronto, Ontario, Canada; Karan Singh, karan@dgp.toronto.edu, University of Toronto, Toronto, Ontario, Canada; Luiz Gustavo Hafemann, luiz.gustavo-hafemann@ubisoft.com, Ubisoft La Forge, Montreal, Quebec, Canada.

SIGGRAPH Conference Papers '25, August 10–14, 2025, Vancouver, BC, Canada

ACM ISBN 979-8-4007-1540-2/2025/08
https://doi.org/10.1145/3721238.3730672

1 Introduction

Automatically generating realistic 3D facial animations from speech has long been a compelling yet challenging goal in computer graphics, with applications in film, virtual reality, video games, and education.

Recent deep generative methods produce realistic facial movements with accurate lip-sync [Sun et al. 2024; Xing et al. 2023; Zhao et al. 2024]. These methods use datasets of paired audio and sequences of 3D meshes, and learn to generate a sequence of facial movements that match the speech. While good lip synchronization may suffice for some applications, *how* the speech is conveyed is

equally important for generating compelling facial performances. There are many ways the same line can be delivered, as many facial movements are only weakly correlated with audio (e.g. movements of eyebrows, forehead, range of the mouth articulation). Providing an intuitive method for *controlling* the motion generation remains an open problem.

Some recent work attempts to add control signals for the generation, by using emotion labels [Daněček et al. 2023] and text [Zhao et al. 2024]. In practice, speech performance is influenced by myriad factors (e.g., character quirks, emotional tone, or speaker identity), making it difficult to annotate or discretize these dimensions without imposing rigid categories. A more flexible approach is to condition the generation based on examples, as investigated in domains such as speech generation [Wang et al. 2018; Zaïdi et al. 2022] and co-speech gesture generation [Ghorbani et al. 2022].

[Sun et al. 2024] provides a first attempt at adding example-based control for speech-driven facial animation. In their work, style is learned in a separate step, with a contrastive loss. However, their approach has key limitations. By using a contrastive loss in a separate training step, the style encoding is learned independently from the main generation task. This means the style encoder cannot learn which facial movements are determined by speech content and which are stylistic choices that vary between performances. Moreover, the contrastive learning framework treats all non-matching examples as equally dissimilar negatives, failing to capture the natural hierarchy of stylistic similarities - for instance, two *angry* performances would be pushed as far apart as an *angry* and a *happy* performance.

We propose a new *example-based* approach for style control that addresses these limitations. Our method takes an audio signal and a style reference, and generates facial animation matching both. The core insight underlying our method is that expressive speech motion can be decomposed into two complementary components: 1) *Time-invariant expressions*, which capture a speaker's overarching expressions throughout line delivery (e.g., squinting eyes or furrowed brows), and 2) *Motion dynamics*, which corresponds to the movements directly tied to phonetic content, capturing how a speaker articulates words. We incorporate this decomposition into a *latent diffusion* model, where the time-invariant expressions modulate the motion dynamics iteratively throughout the diffusion process. By jointly training the style encoder with the motion diffusion model, our method is better suited to extract the nuances of the style reference, and use it to generate facial motion that matches both the style and audio content.

We conducted experiments using a large-scale dataset of paired audio and facial motion data. We compare our model to the state-of-the-art stylization model of [Sun et al. 2024] using a strict experimental protocol, where only the style encoding mechanism is varied. We measured standard metrics on lip synchronization and upper-face movements. We also conducted two user studies to measure the ability of the models in maintaining the style of reference clips, and obtaining accurate lip sync. Our model shows improvement in lip synchronization, upper-face dynamics and style transfer, both in quantitative tests and user studies.

In summary, our key contributions include:

- An example-based speech-driven facial animation framework that captures the speech mannerisms of the reference style, with robust user studies validating its effectiveness in accurately preserving style nuances.
- A novel conditioning mechanism for diffusion-based facial animation generation, called style basis, experimentally shown to improve motion accuracy and expressiveness through ablation studies.
- We release our model's training code and data curation pipeline, to facilitate future research using in-the-wild videos for training and evaluation.

2 Related Works

In these sections, we will discuss related works in three relevant fields, including automatic lip-sync animation, style-conditioned generative models, and latent diffusion models.

2.1 Automatic 3D Lip-sync Animations

The study of automatic 3D lip-sync animation spans several decades, with the correspondence between phonemes (vocalized sounds) and visemes (lip shapes) discovered as early as 1968 [Fisher 1968]. Early lip-sync systems were often procedural, using phoneme alignment to obtain the sequence of phoneme timings, then generate the corresponding sequence of visemes [Cohen and Massaro 1993; Edwards et al. 2016; Massaro et al. 2012; Pan et al. 2022].

More recently, deep learning-based auto-regressive models have been shown to successfully capture 3D lip shapes without explicitly modeling phonemes and visemes [Fan et al. 2022; Karras et al. 2017; Richard et al. 2022; Xing et al. 2023]. However, due to the lack of publicly available 3D speech datasets with rich emotions, these models often focus on relatively neutral speech from the datasets BIWI [Fanelli et al. 2010] and VOCASET [Cudeiro et al. 2019].

Instead of using small datasets of high-quality 3D facial capture, other works build larger-scale datasets from in-the-wild videos. They rely on advancements in monocular facial capture methods [Daněček et al. 2022; Feng et al. 2021; Josi et al. 2024] to extract 3D meshes from videos, commonly modeled as coefficients of a 3D Morphable Model [Egger et al. 2020]. While the quality of the 3D capture from these methods is inferior to that of BIWI and VOCASET, this enables processing large-scale video datasets such as CelebV-text [Yu et al. 2023] and CelebV-HQ [Zhu et al. 2022], and expressive video datasets such as RAVDESS [Livingstone and Russo 2018] and MEAD [Wang et al. 2020]. These datasets have enabled the training of generative models conditioned on emotional categorical labels, such as EMOTE [Daněček et al. 2023] and EmoTalk [Peng et al. 2023], and text-conditioned models such as Media2Face [Zhao et al. 2024].

2.2 Example-based Generative Model

Example-based conditioning is an alternative way to specify the desired style of a generative model. It is commonly used in text-to-speech, since an example audio can intuitively specify a rich set of features at once [Wang et al. 2018; Zaïdi et al. 2022; Zhang et al. 2019]. This approach has also shown promise in co-speech

gesture generation [Ghorbani et al. 2022] and music-conditioned dance generation [Valle-Pérez et al. 2021].

In image-based lip-sync models, an input image can be used not only to establish the appearance, but also the expression and style of the speaker [Xu et al. 2024][Zhang et al. 2023][Zhou et al. 2020]. However, their main challenge relies in maintaining the image integrity rather than copying the desired motion characteristics. For 3D lip-sync models, Imitator [Thambiraja et al. 2022] first learns a motion prior, that produces un-stylized lip-sync motion, to handle stylized speech, it then iteratively optimizes an affine transformation in the latent space to adapt to unseen styles. However, this approach can only adapt viseme shapes while neglecting upper-face expression and global motion trajectories, limiting its applicability to personal viseme generation. This approach also requires model training for each target style.

DiffPoseTalk [Sun et al. 2024] leverages the inherent similarities within motion from the same video clip and the differences across clips through the use of contrastive loss. Using a contrastive loss, they train a motion encoder that compresses a video clip's motion into a fixed-size latent feature, which is used as conditioning for a diffusion model. Nonetheless, the contrastive loss cannot fully capture stylistic similarities among different clips. For example, if two clips depict *angry* speech, the model is still encouraged to produce dissimilar embeddings. As a result, DiffPoseTalk is more suited to learning individualized speech styles from datasets containing long clips of unique speakers, rather than arbitrary-length videos containing different emotional deliveries.

2.3 Latent Diffusion Models

Diffusion models, originally developed for image generation, have become increasingly popular in motion generation, with applications in both speech-driven facial animation [Xu et al. 2024][Zhao et al. 2024] and body animation systems [Karunratanakul et al. 2024][Tevet et al. 2022]. A key advantage of diffusion models is their ability to model many-to-many relationships, effectively avoiding regression to the mean and enabling the generation of diverse and expressive motions [Yang et al. 2024][Tevet et al. 2022]. Latent diffusion models build on this foundation by simplifying the learning process. Instead of operating in the original output space, these models make predictions in a latent space, which reduces complexity and improves efficiency [Rombach et al. 2022].

For 3D facial motion, latent spaces that disentangle expression-related geometry from identity-related geometry are particularly advantageous, as they allow the model to focus on generating motion without having to simultaneously preserve identity features. The FLAME 3D morphable model [Li et al. 2017] is one of the most widely used disentangled latent spaces for facial motion generation and serves as the basis for models like DiffPoseTalk [Sun et al. 2024], Emote [Daněček et al. 2023], and EmoTalk [Peng et al. 2023]. Beyond FLAME, other approaches such as Media2Face [Zhao et al. 2024] and VASA-1 [Xu et al. 2024] have trained their own disentangled feature spaces, demonstrating that the key factor for success lies in disentanglement itself rather than reliance on a specific latent representation. In this work, we learn latent diffusion on the disentangled latent space from SEREP [Josi et al. 2024], as it captures more expressive facial motion compared to FLAME-based methods like [Daněček et al. 2022], being more suitable for evaluating the transfer of talking styles.

3 Design Motivations from User Interviews

To inform our design process for controllable facial animation generation, we conducted interviews with five professional animators at a major game studio company. Each hour-long interview followed a structured protocol covering current workflows, pain points, and reactions to different control paradigms.

All subjects recognized the need for improved tools for automated facial animation, citing an increasing realism gap between generated animations and modern face capture technologies. While subjects appreciated recent approaches such as [Xing et al. 2023] for generating more natural facial movements, they noted limitations in control and emotional expressiveness.

Preferences for controllability clustered around two main use cases: lower-tier animations (e.g., gameplay sequences) and mid-tier animations (scripted sequences). For lower-tier animations, fully automated solutions are preferred, with minimal time spent on control (e.g., providing a tag). For mid-tier animations, animators prefer more control and generation at interactive-time, starting with an initial automated generation that they can then refine with fine-grained controls.

Example-based approaches emerged as a suitable solution for both use cases. For lower-tier animations, a set of pre-defined videos can be used as tags. This offers more flexibility than traditional tag-based approaches, as new tags can be defined without model re-training. For mid-tier animations, these approaches provide finer control by allowing animators to select specific emotional nuances or express hard-to-describe styles through examples.

4 Method

Our proposed model takes as input an audio sequence $\boldsymbol{A}$, an optional style reference $\boldsymbol{X}_{\text{style}} = [x^1, ..., x^M]$, and generates a new motion sequence $\hat{\boldsymbol{X}}_0 = [x^1, ..., x^N]$ such that it lip-syncs to the audio sequence while adhering to the style from $\boldsymbol{X}_{\text{style}}$. To achieve this, we jointly train three key components: (1) a variational autoencoder (VAE) that maps a motion clip into a time-invariant style feature $\boldsymbol{s}$, (2) a style decoder that generates a set of static poses $\boldsymbol{B} = [b^1, \ldots, b^K]$, which we name *style basis*, and (3) a conditional latent diffusion model that predicts the primary motion $\Delta\hat{\boldsymbol{X}}$ alongside a style modulation signal $\boldsymbol{\alpha}$, which governs the influence of the style basis on the primary motion over time. We operate in the latent space of SEREP [Josi et al. 2024] which disentangles speaker identity and expression features. After generating the sequence $\hat{\boldsymbol{X}}_0$, we use the mesh decoder from SEREP to obtain the final mesh animation.

4.1 Conditional Latent Diffusion Formulation

We adopt a latent diffusion model inspired by approaches such as VASA-1 [Xu et al. 2024] and Media2Face [Zhao et al. 2024]. Rather than directly predicting 3D geometry, the model operates in an expression latent space, which constrains the problem and focuses on compact, meaningful representations.

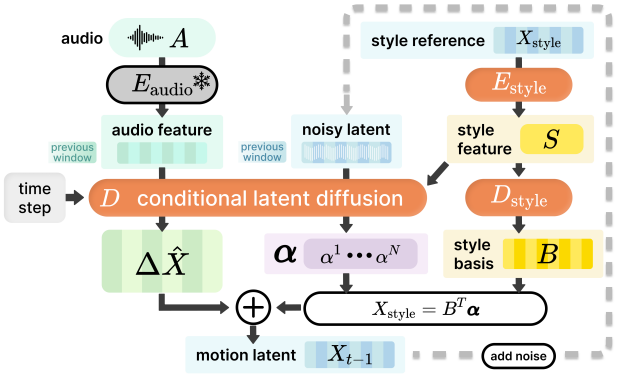

Fig. 2. Overview of the proposed method. An encoder extracts style features S from a style reference clip. These are used, in combination with audio features, to condition a latent diffusion model. The diffusion model outputs the primary motion $\Delta\hat{\boldsymbol{X}}$ and a vector $\boldsymbol{\alpha}$, that modulates the use of the style basis $\boldsymbol{B}$. These are combined to produce the final animation in latent space. This process is repeated for T diffusion steps.

Diffusion models consist of two Markov chains: a forward diffusion process that progressively adds Gaussian noise to a signal $\boldsymbol{X}_0$, resulting in pure noise $\boldsymbol{X}_T$, and a reverse diffusion process $p(\boldsymbol{X}_{t-1} \mid \boldsymbol{X}_t, t, \boldsymbol{C})$ that reconstructs the clean signal through iterative denoising steps [Ho et al. 2020]. The reverse diffusion process is conditioned on a vector $\boldsymbol{C}$ and the diffusion time step t.

To approximate the reverse diffusion process, we train a model $\mathcal{M}(\boldsymbol{X}_t, t, \boldsymbol{C})$. During inference, this model predicts the motion signal from an arbitrary starting noise $\boldsymbol{X}_T$. Similar to prior works [Sun et al. 2024; Tevet et al. 2022; Xu et al. 2024], our model is trained to predict the latent signal and minimize the simple loss $\mathcal{L}_{\text{simple}}$ over the joint distribution of condition C and clean signal $\boldsymbol{X}_0$:

$$\mathbb{E}_{t\sim\mathcal{U}[1,T],\boldsymbol{X}_0,\boldsymbol{C}\sim q(\boldsymbol{X}_0,\boldsymbol{C})}\left[\|\boldsymbol{X}_0 - \mathcal{M}(\boldsymbol{X}_t, t, \boldsymbol{C})\|^2\right]. \tag{1}$$

To accommodate speech of arbitrary length, we use a windowing strategy [Sun et al. 2024; Xu et al. 2024], where the input audio is segmented into fixed-length windows of size N_w. We perform diffusion on each window sequentially. To ensure smooth transitions between windows, we include N_p frames of audio and motion features from previous windows as part of the conditioning signal. We further regularize the motion with velocity loss $\mathcal{L}_{\text{vel}}$ [Cudeiro et al. 2019] and smoothness loss $\mathcal{L}_{\text{smooth}}$ to ensure that the model is learning the correct dynamic and prevents jerky motion:

$$\mathcal{L}_{\text{vel}} = \|(\boldsymbol{X}_{N_p+1:N_w} - \boldsymbol{X}_{N_p:N_w-1}) - (\hat{\boldsymbol{X}}_{N_p+1:N_w} - \hat{\boldsymbol{X}}_{N_p:N_w-1})\|^2, \tag{2}$$

$$\mathcal{L}_{\text{smooth}} = \|\hat{\boldsymbol{X}}_{N_p+2:N_w} - \hat{\boldsymbol{X}}_{N_p+1:N_w-1} + \hat{\boldsymbol{X}}_{N_p:N_w-2}\|^2 \tag{3}$$

4.2 Conditioning Signal

Our conditioning signal combines multiple components: an audio feature $\boldsymbol{A}$, style features $\mathbf{s}$, the last N_p frames of motion latents $\boldsymbol{X}_{\text{prev}}$ from the previous window, and their corresponding audio features $\boldsymbol{A}_{\text{prev}}$. The audio features are produced using a pre-trained HuBERT audio encoder [Hsu et al. 2021], with all layers kept frozen to avoid over-fitting to the training data. Meanwhile, the style feature $\mathbf{s}$ is extracted from an arbitrary-length style reference by the variational auto-encoder E_{style}, which uses a transformer encoder to first generate per-frame style features that considers the varying importance of each frame, then mean-pooled across time steps to ensure it can handle clips of arbitrary length. Within the VAE framework, E_{style} predicts the mean $\boldsymbol{\mu}$ and variance $\boldsymbol{\sigma}^2$ of the latent distribution. The style vector $\mathbf{s}$ is then generated using the re-parameterization trick:

$$\mathbf{s} = \boldsymbol{\mu} + \boldsymbol{\sigma} \odot \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I})$$

This approach allows for differentiable sampling during training. To regularize the latent space, the encoder is optimized with the Kullback-Leibler divergence loss:

$$L_{KL} = \frac{1}{2}\sum_{i=1}^{d}\left(\mu_i^2 + \sigma_i^2 - \log\sigma_i^2 - 1\right)$$

This loss ensures that the learned distribution remains close to a standard normal prior, promoting a smooth and continuous style space. Further implementation details can be found in the supplementary code.

4.3 Guiding Diffusion with a Style Basis

We define *style basis* as: $\boldsymbol{B} = [\boldsymbol{b}_1, \boldsymbol{b}_2, \ldots, \boldsymbol{b}_K]$, where each $\boldsymbol{b}_k \in \mathbb{R}^{1\times D}$ is a latent expression code. Conceptually, these basis vectors capture low-frequency or persistent expressions that are characteristic of the speech mannerism from the style reference. This is inspired by how animators in gaming industry overlay co-speech expressions on top of high frequency lip sync. These are generated from the style latent using the style decoder D_{style},

As shown in Figure 2, we model the diffusion output $\boldsymbol{X} \in \mathbb{R}^{N\times D}$ as the sum of (1) a primary motion $\Delta\hat{\boldsymbol{X}} \in \mathbb{R}^{N\times D}$ and (2) a linear combination of style basis vectors scaled by a time-varying style modulation signal $\boldsymbol{\alpha}_i \in \mathbb{R}^{N\times 1}$:

$$\hat{\boldsymbol{X}} = \Delta\hat{\boldsymbol{X}} + \sum_{i=0}^{K}\boldsymbol{\alpha}_i\,\boldsymbol{b}_i. \tag{4}$$

Each coefficient $\boldsymbol{\alpha}_i$ can vary over time, modulating how much the style basis $\boldsymbol{b}_i$ influences the primary motion throughout the animation. Through the iterative nature of diffusion, the influence of the style basis is incorporated into the primary motion, as shown in Figure 3.

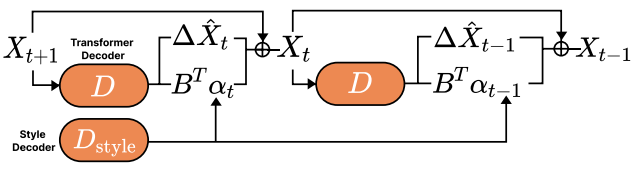

Fig. 3. The style basis $\boldsymbol{B}$ is iteratively incorporated into primary motion throughout the reverse diffusion process, modulated by α_t. Note the reverse diffusion process works in decreasing time step from T to 0.

4.4 Training Strategy

The training objective for our model is composed of the previously specified simple loss and various regularization objectives.

$$\mathcal{L} = \lambda_{\text{simple}}\mathcal{L}_{\text{simple}} + \lambda_{\text{vel}}\mathcal{L}_{\text{vel}} + \lambda_{\text{smooth}}\mathcal{L}_{\text{smooth}} + \lambda_{\text{KL}}\mathcal{L}_{\text{KL}} \quad (5)$$

To train our window-based model, the network must address two scenarios: (a) generating the initial window conditioned on learnable start features, and (b) generating subsequent windows conditioned on speech features and motion parameters from the preceding window. To handle both, we train the model using pairs of examples. Specifically, we sample an audio clip and a motion clip of total length $2N_w$ (zero-padded if shorter) and split them into two windows of length N_w, with the two windows covering scenarios (a) and (b), respectively. In [Sun et al. 2024], it is assumed that the style is consistent across the two windows. Their method employs cross-window conditioning, where window b's motion serves as the style clip to reconstruct window a's motion using window a's audio. However, emotionally expressive speech often involves varying delivery intensities throughout the clip, making this assumption unreliable. To address this, we alternate between cross-conditioning, which uses motion features from the opposite window as the style input, and self-conditioning, which uses motion features from the same window. This strategy mitigates the risk of style leakage which can happen with pure self-conditioning, where the style encoder may encode sequence-specific motion features instead of capturing time-invariant delivery speech styles.

Similar to [Xu et al. 2024], to allow for more nuanced control, and the option to generate un-stylized speech, we also incorporated *classifier free guidance (CFG)* [Ho and Salimans 2022]. During training, we randomly drop 10% audio and style guidance by replacing frames of the corresponding conditioning signal with a learned null condition $\varnothing$. During inference, we apply:

$$\hat{\boldsymbol{X}}_0 = (1 + \sum_{c \in \boldsymbol{C}} \lambda_c) \cdot \mathcal{M}(\boldsymbol{X}_t, t, \boldsymbol{C}) - \sum_{c \in \boldsymbol{C}} \lambda_c \cdot \mathcal{M}(\boldsymbol{X}_t, t, \boldsymbol{C}|_{c=\varnothing}) \quad (6)$$

Where λ_c is the cfg scale for the corresponding condition, and $\boldsymbol{C}|_{c=\varnothing}$ indicates that the corresponding condition is replaced with null condition $\varnothing$.

5 Experimental Protocol

In this section, we detail our experimental protocol, intending to evaluate the impact of different methods for style transfer. For this purpose, we curated a dataset of expressive and stylized speech. To isolate the impact of style transfer methods, we use the same dataset, facial expression representation, overall model architecture and training protocol for all experiments, varying only how style is computed from the style reference, and how it is used for generation.

Dataset. We conducted our experiments using a combination of the CelebV-Text [Yu et al. 2023] and the RAVDESS datasets [Livingstone and Russo 2018]. Both datasets primarily consist of videos featuring a single speaker addressing the camera. The RAVDESS dataset contains videos of the same speakers expressing different emotions, while CelebV-Text consists of unique speakers in each video. Videos in CelebV-Text are typically between 10 to 15 seconds long, whereas RAVDESS videos range from 3 to 6 seconds.

CelebV-Text also includes videos with non-speaking individuals, masked speakers, speakers eating, and speakers facing away from the camera. To filter out these instances, we utilized the dataset's textual tags (e.g., "masked," "eating") and queried only for videos that involved speech, singing, whispering, or reading. Additionally, we used head pose and facial bounding box detection from MediaPipe [Lugaresi et al. 2019] to exclude videos where the speaker's face was angled more than 45 degrees from the camera direction or contained significant head pose jumps. From the remaining videos, we randomly sampled 26.5 hours of content to match the data scale used in prior work, DiffPoseTalk [Sun et al. 2024], which we use for comparison. The entire RAVDESS dataset (3 hours) was used as it does not suffer from video quality issues.

To extract expression coefficients, we first cropped video frames using MediaPipe facial bounding boxes, then applied SEREP [Josi et al. 2024] to compute the facial expression, represented as 64-dimension expression codes, while MediaPipe is used to extract 3-dimension head poses. The full latent representation X at every frame therefore contains 67 dimensions.

For the train-validation split, we randomly selected 90% of the data for training and 10% for validation.

Implementation details. Our model architecture is inspired by DiffPoseTalk [Sun et al. 2024]. The style encoder features a 4-layer transformer encoder (hidden size $d = 512$) with 4 attention heads, followed by 2 fully connected layers with 512 and 256 output features each. The final output is the mean and standard deviation of the style code, each of dimension 128, that are used in the re-parameterization trick for the VAE. For the audio encoder, we utilize the HuBERT model [Hsu et al. 2021] with its feature extractor frozen, followed by a linear layer mapping the output to 512 features.

The motion decoder consists of an 8-layer transformer decoder (with a hidden dimension of 512) with 8 cross-attention heads, attending to both the noisy feature and the audio. The style features are appended to the beginning of the noisy features, allowing the decoder to attend to them at every step. For the style decoder, we use K sets of 3 fully connected layers to map the style features into K style bases, with $K = 4$ for our evaluation experiments. We use a classifier-free guidance strength of 1 ($\lambda_c = 1$) to generate all examples.

We train all modules end-to-end using the Adam optimizer with a learning rate of 0.0001 and a batch size of 16. Data samples are prepared using a window size of $N = 100$ frames with an overlap of $N_p = 10$ frames. The loss function weights are set as follows: $\lambda_{\text{vel}} = 0.5$, $\lambda_{\text{smooth}} = 5 \times 10^{-6}$, and $\lambda_{\text{KL}} = 1 \times 10^{-6}$, experimentally selected to prevent mode collapse during training and kept constant for all experiments. The model is trained on a single NVIDIA A4000 GPU for 450,000 iterations (selected with an early stop mechanism), which takes approximately 4 days.

Baselines. We compare our model against two baselines: the state-of-the-art DiffPoseTalk model [Sun et al. 2024], and a baseline we refer to as VAE-baseline. DiffPoseTalk uses a style-encoder that is pre-trained using a contrastive loss. It then trains a diffusion model conditioned on the style obtained from this pre-trained encoder,

which remains frozen throughout the training process. For VAE-baseline, we implemented a VAE style encoder that is jointly trained with the generative model, in a similar vein to [Ghorbani et al. 2022; Zaïdi et al. 2022]. We employ the same diffusion architecture as the other models for fair comparison, utilizing Denoising Diffusion Probabilistic Models (DDPM) for inference [Ho et al. 2020], with 1000 denoising steps. Our model differs from VAE-baseline by the incorporation of the style basis, and how we employ a mix of cross-window and self-window conditioning. For all evaluations, style references are utilized by all models being compared, ensuring that performance gains are not attributable to any model having access to additional information.

6 Results

Models are evaluated quantitatively using established metrics, qualitatively with examples, and perceptually with a set of user studies.

6.1 Quantitative Results

To quantitatively evaluate the models, we use four established metrics: vertex mean squared error (MSE), lip vertex error (LVE) [Richard et al. 2022], upper-face dynamics deviation (FDD) [Xing et al. 2023], and mouth opening difference (MOD) [Sun et al. 2024]. MSE and LVE measure the per-vertex L_2 error between the target and generated motions, with MSE focusing on the entire face and LVE specifically targeting the vertices around the lips. FDD computes the difference in the standard deviation of each upper-face vertex, serving as a proxy for the dynamics of upper-face motion. MOD evaluates the difference in the extent of the mouth openings, which is indicative of the speech style (e.g. shout vs whisper).

For each test example, we source the style reference from an adjacent but non-overlapping window from the same video as the ground truth. As the clips are adjacent, the style reference contains similar style as the ground truth, ensuring the model has access to the correct style information. By selecting the style clip to be non-overlapping with ground truth, we ensure that we are evaluating style-conditioned generation capabilities rather than motion recreation.

The quantitative results are presented in Table 1. Our method demonstrates improved performance across all metrics compared to baselines, indicating its ability to generate stylistically accurate motion in both the upper and lower face.

The ablation study further shows that most of the performance in LVE and and MSE can be attributed to the inclusion of the style basis, while the inclusion of cross condition allows the model to better capture the dynamics of the upper face, improving FDD.

6.2 Qualitative Results

We refer the reader to the supplementary video (at time 3:44) for video results. Figure 5 compares generated frames from our method and the baselines. These examples are generated using emotionally expressive videos from RAVDESS as the style reference, and audio tracks from CelebV-Text. Notably, this combination lies entirely outside the training distribution, further demonstrating the generalization capability of our method.

Table 1. Quantitative comparison of the proposed model and baselines. Lower values indicate better performance across all metrics. **Bolded** values represent statistically significant gain over all other models ($p < 0.05$).

Model	MSE (mm) ↓	LVE (mm) ↓	FDD ($\times 10^{-5}$m) ↓	MOD (mm) ↓
Unconditioned	4.50	11.0	52.4	3.24
DiffPoseTalk	3.86	10.0	48.3	3.46
VAE-baseline	3.61	9.13	47.4	3.10
MSMD (ours)	**3.46**	**8.54**	**43.4**	**2.96**

Table 2. Quantitative comparison of the proposed model and ablations. Lower values indicate better performance across all metrics. **Bolded** values represent statistically significant gain over all other models ($p < 0.05$).

Model	MSE (mm) ↓	LVE (mm) ↓	FDD ($\times 10^{-5}$m) ↓	MOD (mm) ↓
no style basis	3.61	9.13	47.4	3.10
no cross condition	3.59	8.96	46.3	3.01
MSMD (ours)	**3.46**	**8.54**	**43.4**	2.96

In most cases, DiffPoseTalk either introduces exaggerated expressions or fails to accurately capture the intended style references. This limitation is likely due to its reliance on training the style encoder solely with contrastive loss [Sun et al. 2024], which struggles to capture shared attributes across clips, such as emotions. Similarly, the VAE-baseline model, which excludes the style basis and cross-condition components, produces expressions that roughly align with the reference but frequently omits nuanced upper-face expressions. In contrast, our proposed model effectively captures the reference style, closely matching both the lower-face dynamics and upper-face expressions. We provide additional results of our model rendered with head motion derived from in-the-wild performances, using a 4-second clip as the style reference. See the supplementary video at time 5:15. We further provide visualizations of style basis and example trajectory of alpha (Figure 6, Figure 7 to provide insights to style basis.

6.3 User Study

We conducted user studies to compare the ability of different models on generating animations maintaining the *style* of a style reference, and producing accurate *lip-synchronization*.

We selected 10 videos, considering two scenarios: (a) audio paired with matching style references; (b) audio paired with contrasting style references. For the first scenario, we consider 5 in-the-wild videos, split the videos in two segments, using the start as style reference, and the remaining as the audio input. The second scenario stress-tests the models on the case where the style reference is vastly different than the audio. In this case, we use 5 highly-emotional videos from RAVDESS as style reference, and select in-the-wild audio that do not match the selected emotion for the clips.

To evaluate style transfer and lip synchronization independently, we designed two separate user studies. For each study, participants

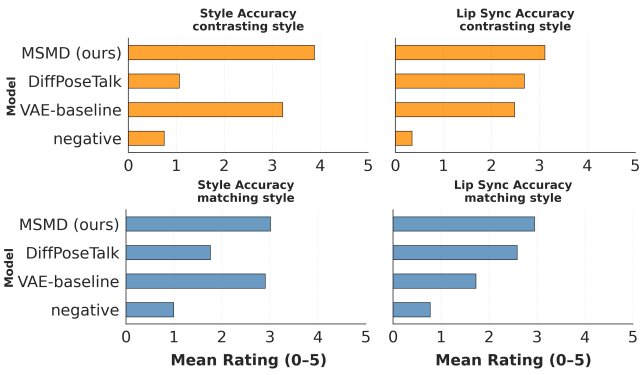

Fig. 4. User-scored performance comparison of the proposed model and baselines for style transfer and lip sync. Results are shown for contrasting style (top) and matching style (bottom), with mean ratings on a 0-5 scale.

were shown the same set of generated animations, rendered without head motion to allow them to focus on facial expressions.

Style transfer evaluation. In the first study, participants evaluated the system's ability to transfer speaking styles from a source video to the generated animation. Participants viewed the style reference video at the top of the screen, followed by four different characters (labeled A, B, C, D). Besides showing the animations from the 3 models under test, we selected a random animation from the validation set (with a different style) as a negative anchor. To focus solely on style transfer quality, videos were presented without audio. Participants rated each character on a scale of 0-5 based on how closely it matched these properties of the style reference video: Facial expressions, Speaking style, and Distinctive facial movements.

Lip synchronization assessment. The lip synchronization study evaluated the temporal alignment between speech and mouth movements. Participants viewed four characters speaking with audio enabled and rated the lip-sync quality on a 0-5 scale, where 0 represented completely mismatched movements and 5 represented perfect synchronization. Besides the 3 models under test, we selected a random animation from the validation set (with different audio) as a negative anchor. To isolate lip-sync quality, participants were instructed to disregard variations in speaking style between characters and focus only on the audio-visual alignment.

Each study was designed to take 10 to 15 minutes. Videos were presented in a looping format to allow participants sufficient time for observation. Participants were encouraged to submit their ratings only when confident in their assessment. This approach allowed for careful evaluation while maintaining reasonable time constraints for the complete study session. Please refer to the supplementary material for the detailed instructions given to participants.

We collected 34 and 32 responses for the style accuracy and lip sync accuracy experiments, respectively. The results of this evaluation are summarized in Figure 4. We observe that our model has superior performance both at capturing the style and generating the correct lip sync, especially for capturing highly expressive speech styles from the RAVDESS examples (contrasting style). A set of paired t-tests confirms that our model's mean ratings are significantly higher than all competing methods ($p < 0.0005$).

On the other hand, DiffPoseTalk was unable to capture the speech styles in most cases, showing the limitations of training the style encoder as a separate task with contrastive loss. Both DiffPoseTalk and the VAE-baseline show worse lip sync capability than the proposed model. We hypothesize that our performance gain is due to the style basis representation, which effectively decouples speech from style, preventing style from disrupting essential viseme shapes.

7 Discussions

7.1 Use Case of Style-conditioned Generation

Style-conditioned generation enables two distinct applications: First, it can seamlessly integrate into existing pipelines that rely on tag-based animations (see Section 3). By curating a set of style references to represent various tags, offering greater flexibility. The curated examples can be easily updated with more suitable ones, or the tag library can be expanded to accommodate new styles.

Secondly, style-conditioned generation can be applied to produce dubbed animations while preserving certain aspects of the original performance. For instance, the original performance can serve as the style reference, while using an input audio track in a different language. This approach is more versatile than simply replacing lip movements, as it allows for dubbing over the original performance with audio of varying lengths, maintaining expressive alignment.

7.2 Limitations

While our model can mimic the speech style of arbitrary style references to any audio while maintaining lip sync coherence, it does not consider whether the delivery style of the speech, such as tone, pace, and duration matches the delivery style in the reference. For example, the model can map an angry speech style to an audio with an uplifting tone which would generate uncanny animation (see the supplementary video at time 4:54).

Furthermore, our model is designed based on the definition that delivery style is something invariant throughout the delivery of a line of speech. In the case where speech transitions from one style to another (such as from a neutral expression to an angry expression), we don't have a way to innately control the transition, and the animator would have to manually break the audio into windows and use different style for each window.

Lastly, our model does not consider replicating isolated instances of iconic gestures such as the eyebrow raise of Dwayne "the rock" Johnson, or eye-rolling. Which can be considered very significant to the style of a performance. A potential future direction could be finding a way to incorporate these facial gestures into generation.

7.3 Future Works

Example-based motion generation offers an intuitive way to control animation; however, finding an exact style reference that matches the desired style can be challenging. For instance, a user may want to extract upper-face motion from one example while utilizing lip motion from another. A promising direction for future work is developing a system that allows finer control over style attributes by enabling users to specify and blend motion characteristics from

multiple style references to better influence the generation process. Another promising direction is to integrate our VAE-based style space with semantically rich representations, such as the CLIP space proposed in [Ma et al. 2024]. This integration would enable multi-modal conditioning, allowing the model to generate facial animations informed by both video and text inputs.

8 Conclusion

We introduced an example-based speech-driven facial animation framework that captures the distinctive motion characteristics and expressive essence of a reference style while maintaining accurate lip synchronization, as demonstrated through user studies. Our proposed conditioning mechanism, style basis, has been shown to enhance motion accuracy compared to its ablated counterpart, validating its contribution to the overall performance. This work highlights the potential of example-conditioned facial animation generation, offering a complementary approach to text-conditioned generation.

Acknowledgments

We thank Arthur Josi for the assistance working with the SEREP latent model, Emeline Got for rendering the beautiful animations, Marc-André Carbonneau for motivating the idea of using example-based conditioning, and Domenico Tullo for the valuable feedback on our user study design. This project is funded by the Mitacs Accelerate program, IT40568.

References

Michael M. Cohen and Dominic W. Massaro. 1993. Modeling Coarticulation in Synthetic Visual Speech. In *Models and Techniques in Computer Animation*, Nadia Magnenat Thalmann and Daniel Thalmann (Eds.). Springer Japan, Tokyo, 139–156. https://doi.org/10.1007/978-4-431-66911-1_13

Daniel Cudeiro, Timo Bolkart, Cassidy Laidlaw, Anurag Ranjan, and Michael J. Black. 2019. Capture, Learning, and Synthesis of 3D Speaking Styles. https://doi.org/10.48550/arXiv.1905.03079 arXiv:1905.03079 [cs] version: 1.

Radek Daněček, Kiran Chhatre, Shashank Tripathi, Yandong Wen, Michael Black, and Timo Bolkart. 2023. Emotional speech-driven animation with content-emotion disentanglement. In *SIGGRAPH Asia 2023 Conference Papers*. 1–13.

Radek Daněček, Michael J. Black, and Timo Bolkart. 2022. EMOCA: Emotion Driven Monocular Face Capture and Animation. https://doi.org/10.48550/arXiv.2204.11312 arXiv:2204.11312 [cs].

Radek Daněček, Kiran Chhatre, Shashank Tripathi, Yandong Wen, Michael J. Black, and Timo Bolkart. 2023. Emotional Speech-Driven Animation with Content-Emotion Disentanglement. In *SIGGRAPH Asia 2023 Conference Papers*. 1–13. https://doi.org/10.1145/3610548.3618183 arXiv:2306.08990 [cs].

Pif Edwards, Chris Landreth, Eugene Fiume, and Karan Singh. 2016. JALI: an animator-centric viseme model for expressive lip synchronization. *ACM Transactions on Graphics* 35, 4 (July 2016), 1–11. https://doi.org/10.1145/2897824.2925984

Bernhard Egger, William A. P. Smith, Ayush Tewari, Stefanie Wuhrer, Michael Zollhoefer, Thabo Beeler, Florian Bernard, Timo Bolkart, Adam Kortylewski, Sami Romdhani, Christian Theobalt, Volker Blanz, and Thomas Vetter. 2020. 3D Morphable Face Models—Past, Present, and Future. *ACM Transactions on Graphics* 39, 5 (June 2020), 157:1–157:38. https://doi.org/10.1145/3395208

Yingruo Fan, Zhaojiang Lin, Jun Saito, Wenping Wang, and Taku Komura. 2022. FaceFormer: Speech-Driven 3D Facial Animation with Transformers. https://doi.org/10.48550/arXiv.2112.05329 arXiv:2112.05329 [cs].

Gabriele Fanelli, Juergen Gall, Harald Romsdorfer, Thibaut Weise, and Luc Van Gool. 2010. A 3-D Audio-Visual Corpus of Affective Communication. *IEEE Transactions on Multimedia* 12, 6 (Oct. 2010), 591–598. https://doi.org/10.1109/TMM.2010.2052239 Conference Name: IEEE Transactions on Multimedia.

Yao Feng, Haiwen Feng, Michael J. Black, and Timo Bolkart. 2021. Learning an Animatable Detailed 3D Face Model from In-The-Wild Images. https://doi.org/10.48550/arXiv.2012.04012 arXiv:2012.04012 [cs].

Cletus G. Fisher. 1968. Confusions Among Visually Perceived Consonants. *Journal of Speech and Hearing Research* 11, 4 (Dec. 1968), 796–804. https://doi.org/10.1044/jshr.1104.796 Publisher: American Speech-Language-Hearing Association.

Saeed Ghorbani, Ylva Ferstl, Daniel Holden, Nikolaus F. Troje, and Marc-André Carbonneau. 2022. ZeroEGGS: Zero-shot Example-based Gesture Generation from Speech. https://doi.org/10.48550/arXiv.2209.07556 arXiv:2209.07556 [cs].

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising Diffusion Probabilistic Models. https://doi.org/10.48550/arXiv.2006.11239 arXiv:2006.11239 [cs].

Jonathan Ho and Tim Salimans. 2022. Classifier-Free Diffusion Guidance. https://doi.org/10.48550/arXiv.2207.12598 arXiv:2207.12598 [cs].

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units. https://doi.org/10.48550/arXiv.2106.07447 arXiv:2106.07447 [cs].

Arthur Josi, Luiz Gustavo Hafemann, Abdallah Dib, Emeline Got, Rafael M. O. Cruz, and Marc-Andre Carbonneau. 2024. SEREP: Semantic Facial Expression Representation for Robust In-the-Wild Capture and Retargeting. https://doi.org/10.48550/arXiv.2412.14371 arXiv:2412.14371 [cs].

Tero Karras, Timo Aila, Samuli Laine, Antti Herva, and Jaakko Lehtinen. 2017. Audio-driven facial animation by joint end-to-end learning of pose and emotion. *ACM Trans. Graph.* 36, 4 (July 2017), 94:1–94:12. https://doi.org/10.1145/3072959.3073658

Korrawe Karunratanakul, Konpat Preechakul, Emre Aksan, Thabo Beeler, Supasorn Suwajanakorn, and Siyu Tang. 2024. Optimizing Diffusion Noise Can Serve As Universal Motion Priors. https://doi.org/10.48550/arXiv.2312.11994 arXiv:2312.11994 [cs].

Tianye Li, Timo Bolkart, Michael J. Black, Hao Li, and Javier Romero. 2017. Learning a model of facial shape and expression from 4D scans. *ACM Transactions on Graphics* 36, 6 (Dec. 2017), 1–17. https://doi.org/10.1145/3130800.3130813

Steven R. Livingstone and Frank A. Russo. 2018. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. *PLOS ONE* 13, 5 (May 2018), e0196391. https://doi.org/10.1371/journal.pone.0196391 Publisher: Public Library of Science.

Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, Wan-Teh Chang, Wei Hua, Manfred Georg, and Matthias Grundmann. 2019. MediaPipe: A Framework for Building Perception Pipelines. https://doi.org/10.48550/arXiv.1906.08172 arXiv:1906.08172 [cs].

Yifeng Ma, Suzhen Wang, Yu Ding, Bowen Ma, Tangjie Lv, Changjie Fan, Zhipeng Hu, Zhidong Deng, and Xin Yu. 2024. TalkCLIP: Talking Head Generation with Text-Guided Expressive Speaking Styles. https://doi.org/10.48550/arXiv.2304.00334 arXiv:2304.00334 [cs].

D. W. Massaro, M. M. Cohen, M. Tabain, J. Beskow, and R. Clark. 2012. Animated speech: research progress and applications. In *Audiovisual Speech Processing* (1 ed.), Gerard Bailly, Pascal Perrier, and Eric Vatikiotis-Bateson (Eds.). Cambridge University Press, 309–345. https://doi.org/10.1017/CBO9780511843891.014

Yifang Pan, Chris Landreth, Eugene Fiume, and Karan Singh. 2022. VOCAL: Vowel and Consonant Layering for Expressive Animator-Centric Singing Animation. In *SIGGRAPH Asia 2022 Conference Papers (SA '22)*. Association for Computing Machinery, New York, NY, USA, 1–9. https://doi.org/10.1145/3550469.3555408

Ziqiao Peng, Haoyu Wu, Zhenbo Song, Hao Xu, Xiangyu Zhu, Jun He, Hongyan Liu, and Zhaoxin Fan. 2023. EmoTalk: Speech-Driven Emotional Disentanglement for 3D Face Animation. https://doi.org/10.48550/arXiv.2303.11089 arXiv:2303.11089 [cs].

Alexander Richard, Michael Zollhoefer, Yandong Wen, Fernando de la Torre, and Yaser Sheikh. 2022. MeshTalk: 3D Face Animation from Speech using Cross-Modality Disentanglement. https://doi.org/10.48550/arXiv.2104.08223 arXiv:2104.08223 [cs].

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-Resolution Image Synthesis with Latent Diffusion Models. https://doi.org/10.48550/arXiv.2112.10752 arXiv:2112.10752 [cs].

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

Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H. Bermano. 2022. Human Motion Diffusion Model. https://doi.org/10.48550/arXiv.2209.14916 arXiv:2209.14916 [cs].

Balamurugan Thambiraja, Ikhsanul Habibie, Sadegh Aliakbarian, Darren Cosker, Christian Theobalt, and Justus Thies. 2022. Imitator: Personalized Speech-driven 3D Facial Animation. https://doi.org/10.48550/arXiv.2301.00023 arXiv:2301.00023 [cs].

Guillermo Valle-Pérez, Gustav Eje Henter, Jonas Beskow, André Holzapfel, Pierre-Yves Oudeyer, and Simon Alexanderson. 2021. Transflower: probabilistic autoregressive dance generation with multimodal attention. *ACM Transactions on Graphics* 40, 6 (Dec. 2021), 1–14. https://doi.org/10.1145/3478513.3480570 arXiv:2106.13871 [cs].

Kaisiyuan Wang, Qianyi Wu, Linsen Song, Zhuoqian Yang, Wayne Wu, Chen Qian, Ran He, Yu Qiao, and Chen Change Loy. 2020. MEAD: A Large-Scale Audio-Visual Dataset for Emotional Talking-Face Generation. In *Computer Vision – ECCV 2020*, Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm (Eds.). Vol. 12366. Springer International Publishing, Cham, 700–717. https://doi.org/10.1007/978-3-030-58589-1_42 Series Title: Lecture Notes in Computer Science.

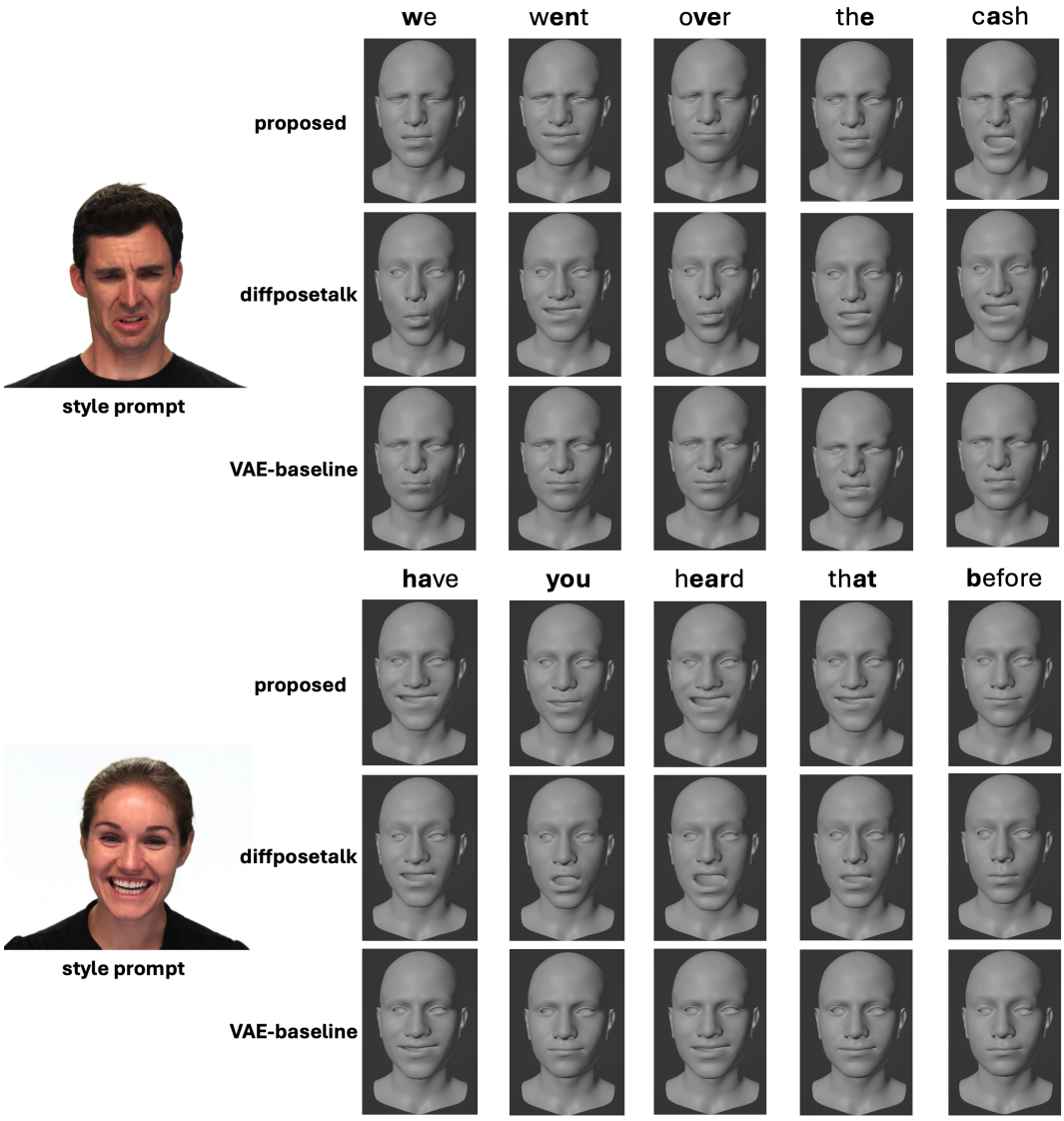

Fig. 5. Qualitative results of baselines and proposed models on different style prompts and audio.

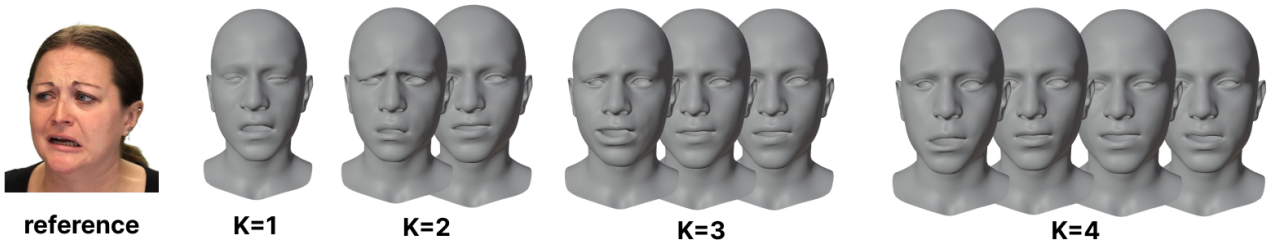

Fig. 6. Example of style basis from models of basis size 1, 2, 3, and 4 and the corresponding reference.

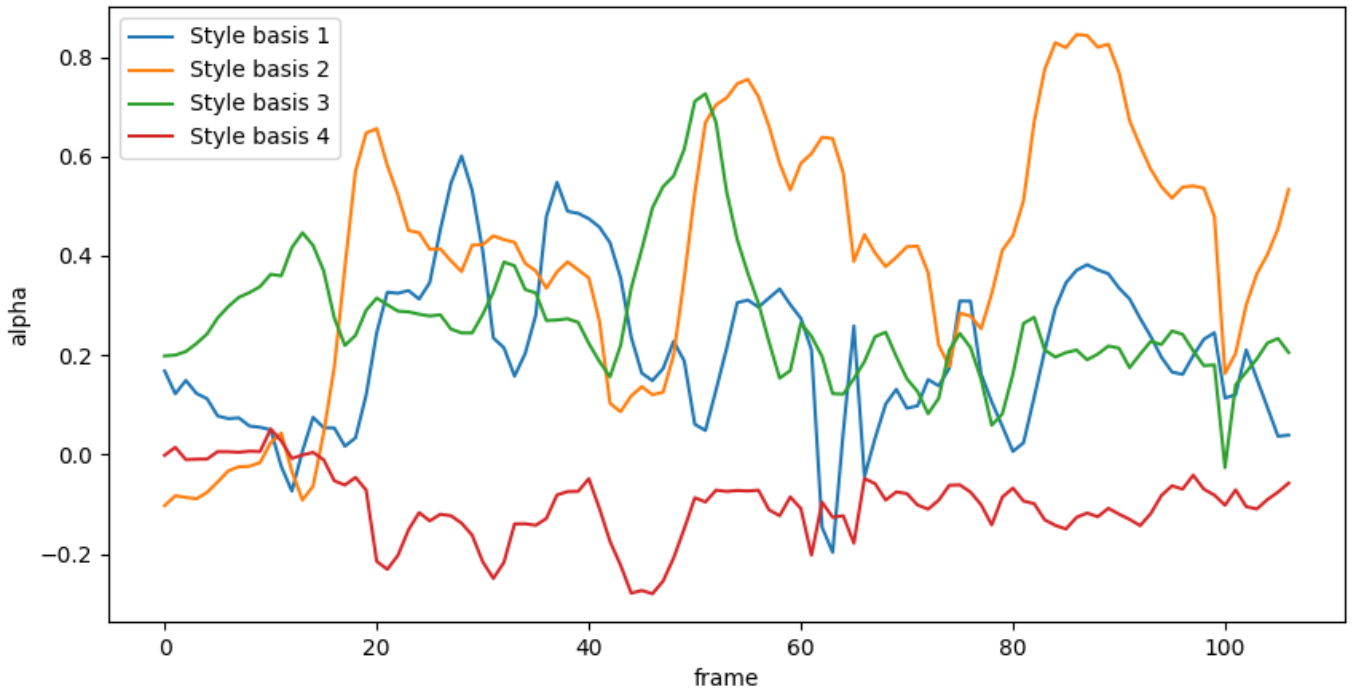

Fig. 7. Example of alpha trajectory over time for 4 style basis.

Yuxuan Wang, Daisy Stanton, Yu Zhang, R. J. Skerry-Ryan, Eric Battenberg, Joel Shor, Ying Xiao, Fei Ren, Ye Jia, and Rif A. Saurous. 2018. Style Tokens: Unsupervised Style Modeling, Control and Transfer in End-to-End Speech Synthesis. https://doi.org/10.48550/arXiv.1803.09017 arXiv:1803.09017 [cs].

Jinbo Xing, Menghan Xia, Yuechen Zhang, Xiaodong Cun, Jue Wang, and Tien-Tsin Wong. 2023. CodeTalker: Speech-Driven 3D Facial Animation with Discrete Motion Prior. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Vancouver, BC, Canada, 12780–12790. https://doi.org/10.1109/CVPR52729.2023.01229

Sicheng Xu, Guojun Chen, Yu-Xiao Guo, Jiaolong Yang, Chong Li, Zhenyu Zang, Yizhong Zhang, Xin Tong, and Baining Guo. 2024. VASA-1: Lifelike Audio-Driven Talking Faces Generated in Real Time. https://doi.org/10.48550/arXiv.2404.10667 arXiv:2404.10667 [cs].

Hongdi Yang, Chengyang Li, Zhenxuan Wu, Gaozheng Li, Jingya Wang, Jingyi Yu, Zhuo Su, and Lan Xu. 2024. SMGDiff: Soccer Motion Generation using diffusion probabilistic models. https://doi.org/10.48550/arXiv.2411.16216 arXiv:2411.16216 [cs].

Jianhui Yu, Hao Zhu, Liming Jiang, Chen Change Loy, Weidong Cai, and Wayne Wu. 2023. CelebV-Text: A Large-Scale Facial Text-Video Dataset. https://doi.org/10.48550/arXiv.2303.14717 arXiv:2303.14717 [cs].

Julian Zaïdi, Hugo Seuté, Benjamin van Niekerk, and Marc-André Carbonneau. 2022. Daft-Exprt: Cross-Speaker Prosody Transfer on Any Text for Expressive Speech Synthesis. In *Interspeech 2022*. 4591–4595. https://doi.org/10.21437/Interspeech.2022-10761 arXiv:2108.02271 [cs].

Wenxuan Zhang, Xiaodong Cun, Xuan Wang, Yong Zhang, Xi Shen, Yu Guo, Ying Shan, and Fei Wang. 2023. SadTalker: Learning Realistic 3D Motion Coefficients for Stylized Audio-Driven Single Image Talking Face Animation. https://doi.org/10.48550/arXiv.2211.12194 arXiv:2211.12194 [cs].

Ya-Jie Zhang, Shifeng Pan, Lei He, and Zhen-Hua Ling. 2019. Learning Latent Representations for Style Control and Transfer in End-to-end Speech Synthesis. https://doi.org/10.48550/arXiv.1812.04342 arXiv:1812.04342 [cs].

Qingcheng Zhao, Pengyu Long, Qixuan Zhang, Dafei Qin, Han Liang, Longwen Zhang, Yingliang Zhang, Jingyi Yu, and Lan Xu. 2024. Media2Face: Co-speech Facial Animation Generation With Multi-Modality Guidance. https://doi.org/10.48550/arXiv.2401.15687 arXiv:2401.15687 [cs].

Yang Zhou, Xintong Han, Eli Shechtman, Jose Echevarria, Evangelos Kalogerakis, and Dingzeyu Li. 2020. MakeItTalk: Speaker-Aware Talking-Head Animation. *ACM Transactions on Graphics* 39, 6 (Dec. 2020), 1–15. https://doi.org/10.1145/3414685.3417774 arXiv:2004.12992 [cs].

Hao Zhu, Wayne Wu, Wentao Zhu, Liming Jiang, Siwei Tang, Li Zhang, Ziwei Liu, and Chen Change Loy. 2022. CelebV-HQ: A Large-Scale Video Facial Attributes Dataset. https://doi.org/10.48550/arXiv.2207.12393 arXiv:2207.12393 [cs].