

Let's Chorus: Partner-aware Hybrid Song-Driven 3D Head Animation

Xiumei Xie¹ Zikai Huang¹ Wenhao Xu¹ Peng Xiao¹ Xuemiao Xu^{1,2†} Huaidong Zhang^{1,2†}
¹South China University of Technology
²Guangdong Engineering Center for Large Model and GenAI Technology

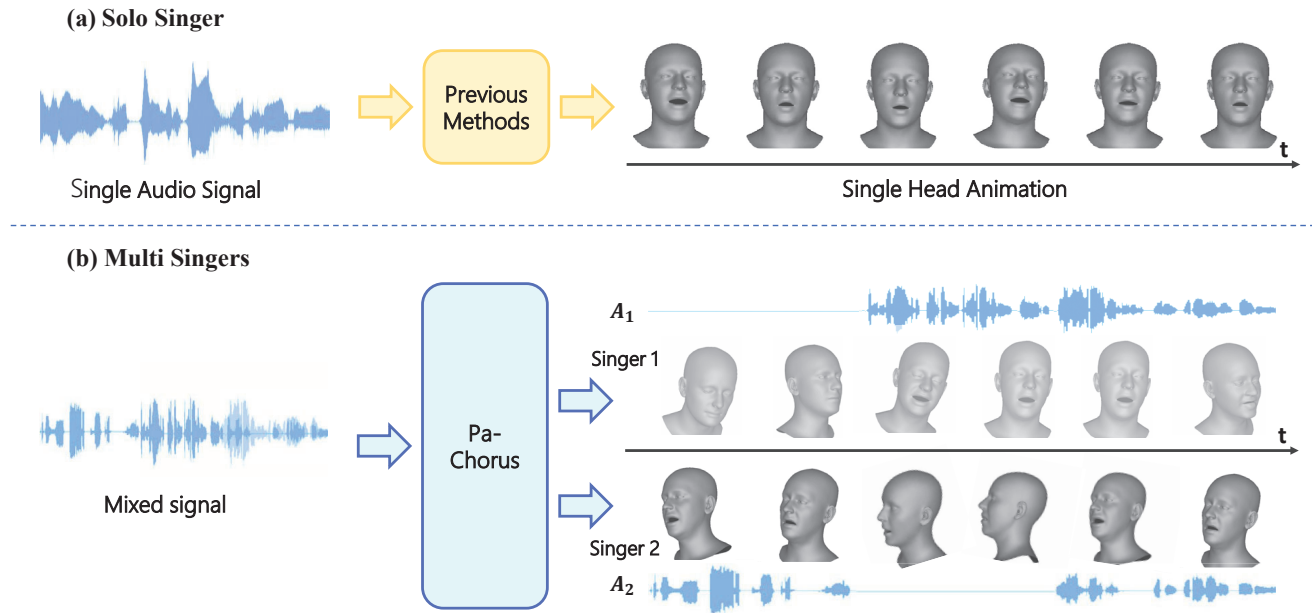


Figure 1. Multi-singers Animation. (a): Previous methods construct the 3D facial animation conditioned with an input of single-person audio. (b): With a hybrid song from multi-singers, we argue that it is essential to construct the emotional interaction between each singer for accurate 3D head generation. Motivated by this, we propose the *PaChorus* framework conditioned on a segment of mixed audio consisting of background music and vocals from multi-singers. With inter-singer interaction modeling, our method can generate emotion-consistent animation sequences.

Abstract

Singing is a vital form of human emotional expression and social interaction, distinguished from speech by its richer emotional nuances and freer expressive style. Thus, investigating 3D facial animation driven by singing holds significant research value. Our work focuses on 3D singing facial animation driven by mixed singing audio, and to the best of our knowledge, no prior studies have explored this area. Additionally, the absence of existing 3D singing datasets poses a considerable challenge. To address this, we collect a novel audiovisual dataset, ChorusHead which features synchronized mixed vocal audio and 3D motions for chorus singing. In addition, We propose a partner-aware

3D chorus head generation framework driven by mixed audio inputs. The proposed framework extracts emotional features from the background music and dependence between singers and models the head movement in a latent space from the Variational Autoencoder (VAE), enabling diverse interactive head animation generation. Extensive experimental results demonstrate that our approach effectively generates 3D facial animations of interacting singers, achieving notable improvements in realism and handling background music interference with strong robustness. The dataset will be released for research purposes at the project page: <https://xxiexm.github.io/PaChorus/>.

[†]Corresponding author (xuexam@scut.edu.cn, huaidongz@scut.edu.cn).

1. Introduction

Animating facial movements to reflect singing performances has significant applications in virtual entertainment, digital avatars, and immersive user experiences. Compared to conversational facial animation, singing-driven facial animation requires a higher level of expressiveness and synchronization with complex audio inputs. Unlike speech, which often involves limited head and lip movements, singing incorporates broader head gestures, dynamic changes in mouth shape, and alignment with rhythm and melody, making it a more challenging task. Capturing these nuances accurately is essential for producing lifelike and engaging singing animations.

Despite recent advances in 3D facial animation, the majority of approaches [1–4], have focused predominantly on speech-driven animation. Traditional methods typically rely on speech audio-to-lip mapping or facial keypoint generation. Many of these methods utilize pre-defined facial expressions or employ frame-by-frame synthesis that captures basic lip and jaw movements. However, such approaches are limited in their scope and cannot be easily adapted to singing-driven animation due to their lack of emphasis on musical expressiveness and synchronization with rhythm and melody. More recently, the generation of audio-driven animation has attracted more attention in the community [5–8]. However, these methods construct the pipeline with the condition of a mono-singer’s audio. Despite the generalization of these methods to multi-singer scenarios is possible, the modeling of interaction features between singers is out of consideration, as shown in Fig. 1 (a). These limitations hinder their effectiveness in achieving the naturalistic head movements, interactive gestures, and tempo alignment crucial for realistic chorus-head animation.

To address the above challenges, we introduce a novel task: singing head animation driven by mixed singing voices. We present the *PaChorus*, *Partner-aware Chorus* 3D head generation framework, which is designed to learn the emotional features from the background music and interaction between singers, enabling the realistic generation of lip, facial, and head 3D animation. As shown in Fig. 1 (b), when Singer1 or Singer2 is silent, their mouths remain closed, and only the head moves subtly in sync with the background music. During segments where both singers are vocalizing, interactive gestures and synchronized singing become more evident, enhancing the sense of chorus performance.

Our framework is specifically designed to handle complex singing input from multiple vocal sources, synthesizing facial animations that capture not only accurate lip-sync but also expressive head and upper facial movements synchronized to the musical beat. By leveraging advancements in multi-speaker separation [9–12] and rhythm-aware gesture synthesis [13–16], our model bridges the gap be-

tween traditional facial animation and the unique demands of singing, offering a new pathway toward realistic, expressive singing animations.

In summary, our main contributions are as follows:

- **First exploration of multi-singer facial animation driven by mixed vocals:** This work pioneers the study of generating expressive 3D facial animations for multi-singer choral performances driven by mixed audio, setting a foundation for future research.
- **Introduction of the Chorus Dataset:** We propose the first 3D multi-singer, multi-modal singing dataset, *ChorusHead*, which extends the research potential for downstream tasks in this domain.
- **Latent space for interactive head pose generation:** We design a VAE-based latent space with stochastic combinations conditioned on head pose generation in the proposed PaChorus framework, to improve interaction and realism in multi-singer animation.

We also conduct various comparisons on the ChorusHead dataset with the existing methods. Extensive experiments show that the proposed method can generate more realistic head animation in the condition of mixture audio.

2. Related Work

Extensive research has delved into the domain of speech-driven 3D facial animation generation, yet the landscape of song-driven animation remains largely unexplored. This is primarily due to the scarcity of singing-specific datasets and the distinctive acoustic and aesthetic complexities inherent to singing. In light of this, we present a comprehensive review of the latest advancements in speech-driven 3D facial animation, alongside an examination of emerging efforts that focus on song-driven animation for a solo singer.

2.1. Speech-Driven Facial Animation Synthesis

The generation of audio-driven facial animations has become a significant focus within the field of multi-modal visual synthesis, with broad applications in modern multimedia, including virtual avatars, gaming, and film production. Research in this domain can be classified into two primary categories: 2D-based and 3D-based methods. 2D approaches often rely on techniques like optical flow [17], keypoint detection [18], or disentangled representations [19] to create lifelike facial animations synchronized with audio. While these methods have seen success in applications such as film dubbing and 2D gaming, they face notable limitations in extending to 3D applications, such as VR/AR, film production, and 3D gaming, where the complexity of spatial dynamics demands more robust solutions.

To address the need for 3D facial animation, several rule-based methods have been proposed [20–22], aiming to map input audio to 3D facial rigs. However, these rule-based systems typically require labor-intensive manual annotations

and fine-tuning of parameters to achieve realistic results. To mitigate this, data-driven approaches have emerged [23–26], which leverage large datasets to automate the learning of audio-to-facial mappings. Some methods [25] utilize monocular 3D face reconstruction from video data, demonstrating strong generalization across different subjects, but the limitations of monocular reconstruction techniques can restrict the quality of the final animation.

VOCA [27] employed CNN to map audio to facial expressions. FaceFormer [1], the first approach to introduce Transformer architectures in this domain, models both the alignment and contextual relationships between audio and 3D vertices, achieving notable performance. Subsequent Transformer-based methods [2–4] have further demonstrated reliable facial animation for speech-driven scenarios. Recently, diffusion-based methods [28–30] have gained attention, with works such as DiffPosetalk [30] and FaceDiffuser [28] improving the diversity of output while maintaining the quality of the generation. To improve emotional expressiveness, Emotalk [31] and EMOTE [32] propose a decoupling framework that separates emotional content from linguistic information through emotional exchange and cross-reconstruction mechanisms. This strategy effectively enhances the emotional representation capacity of facial animation systems. In parallel, ExpClip [33] introduces a prompt-based approach to annotate facial action units (AUs), enabling finer-grained control over facial expressions.

Despite these advancements, existing methods primarily focus on individual speakers and speech-driven contexts, often overlooking the complexities of interpersonal dynamics, which is essential for creating highly expressive and interactive animations. Moreover, while speech typically involves constrained expressions, singing, particularly in multi-singer choir performances, demands more exaggerated emotions and a wider range of head and facial movements to convey expressive intent. Therefore, addressing these challenges necessitates the development of a dedicated dataset focused on multi-singer choir interactions, which is crucial for advancing research in this domain.

2.2. Song-Driven Solo Facial Animation

In recent years, researchers have made initial attempts in the field of singing-driven facial animation generation, both in 2D and 3D contexts [5], [6], [7]. VOCAL [5] developed the Ma-Ps singing model based on rich domain priors, focusing primarily on lower facial predictions, while neglecting the upper face and head pose, which are crucial for expressive singing. Song2Face [6] proposed a network driven by singing voice and singer labels to generate facial animations, yet this approach is limited to handling vocals without background music interference. However, in typical musi-

cal audio signals, vocals are often intertwined with background music. To address this challenge, MusicFace [7] developed a decoupling and fusion strategy that separates audio signals into vocals and background music, leveraging an attention mechanism to model their interaction. Singing-Head [8] introduced a 27-hour 3D singing dataset, but the captured performances are limited to individual singers.

It is noteworthy that the nature of chorus singing, characterized by heightened expressiveness and complex interactions among multiple singers, poses unique challenges not addressed by current methodologies. To fill this gap, we develop a multimodal dataset featuring diverse and expressive 3D singing animations derived from aligned audio-text-flame triplets, aimed at creating vibrant singing representations that align with both content and emotional expression. This new dataset will support multilingual capabilities and a broad thematic range, thus addressing the limitations of current single-singer approaches and paving the way for more sophisticated multi-singer animation generation.

3. ChorusHead Dataset

Considering the absence of the chorus 3D singing facial animation dataset, we present ChorusHead, a large-scale dataset with chorus audio and motion. The dataset comparison is shown in Tab. 1. Specifically, we collected a large dataset of publicly available chorus singing videos from the internet, with a resolution of 1024×1024 at 30 FPS. To provide a dataset with reliable labeling for chorus model training and testing, we process the collected data in three steps: audio track separation, video clipping and filtering, and 3D head animation generation.

Audio track separation. To collect the vocal track for each singer, we pass the mixed vocal audio through a song separation model to obtain distinct audio tracks. We tested various state-of-the-art pre-trained models [9–11, 34] in the field of audio processing on our dataset and observed minimal performance differences between audio-visual and audio-only methods. To achieve a balance between model efficiency and separation quality, we opted for an audio-only approach, incorporating MossFormer2 [35] as the tool for voice separation. Notably, nearly all existing models experience significant performance degradation in noisy environments, highlighting a key unresolved issue in voice separation research. To address this, we utilized spleeter [36] to separate voice and background music, effectively simplifying the task and producing cleaner vocal output.

Video clipping and filtering. For the initial filtering of raw videos, we prioritized selecting chorus videos where singers face forward with a fixed camera, which makes it easier to generate reliable 3D head animation. Also, since the current version of MossFormer2 [35] only performs well on duet song separation, we only reserve the video with two singers for further labeling. A streamlined and effective au-

tomated workflow was designed as follows: First, we applied Spleeter [36] and Voice activity detection(VAD) [37] to detect the start and end timestamps of the vocal segments, trimming any unnecessary silent sections from the videos. Next, we split the video frames into two halves, labeled as Singer1 and Singer2, respectively. Using MTCNN [38], we then cropped the faces from both halves and saved them, aligning the image indices with the video frame numbers. Since some frames may have missing faces in either half, we performed a second filtering step by identifying contiguous sequences longer than 150 frames (5 seconds at 30 FPS) in both Singer1 and Singer2. We also annotated the range where faces are simultaneously present in both parts.

3D head animation generation. To ensure the quality of facial expressions and mouth movements, we used EMOCA v2 [39] to extract pseudo-GT, consisting of 53 expression (face + jaw) coefficients and 3 pose coefficients. To address jitter issues caused by certain factors, we applied Gaussian smoothing. In the final filtering step, we visualized the pseudo-GT alongside the corresponding RGB video, manually removing segments with minimal mouth movement or evident errors.

Table 1. Comparison of talking/singing face animation datasets.

Datasets	Task	Dura.	BGM	2D	3D	Multi-person	Inter-head
MEAD [40]	Talking	40.0h	-	✓	-	-	-
HDTF [41]	Talking	15.8h	-	✓	-	-	-
VOCASET [27]	Talking	0.5h	-	-	✓	-	-
3D-ETF [31]	Talking	6.5h	-	-	✓	-	-
RAVDESS [42]	Singing	2.6h	-	✓	-	-	-
Song2Face [6]	Singing	2.0h	✓	✓	-	-	-
Musicface [7]	Singing	40.0h	✓	✓	-	-	-
SingingHead [8]	Singing	27.0h	✓	✓	✓	-	-
ChorusHead (Ours)	Singing	8.0h	✓	-	✓	✓	✓

4. Methodology

Due to the limitation of the song separation model MossFormer2 [35], we only utilize the videos with two singers for model training. For clarity in this section, we formulate the problem definition and present our method in the scope of two singers. Kindly note that the proposed trained model is flexible for three or more singers’ chorus animation generation. Please refer to Sec. 5.5 for more details.

4.1. Problem Definition

In the case of two singers, given a choral audio \mathcal{M} accompanied by background music, our goal is to generate expressive animations $\mathcal{F} = (F^{S1}, F^{S2})$ for multiple singers while also modeling the inter-singers interaction. Each \mathcal{M} consists of two vocal sequences $\mathcal{A}_{1:t}^{S1}, \mathcal{A}_{1:t}^{S2}$ and background melody sequence $\mathcal{A}_{1:t}^B$, where $t \in [1, T]$ indicates the number of the frames. For each singer, we use $F^S = (f_1^S, \dots, f_t^S)$ to represent 3D motion sequence. Specifically, we represent any motions using a 3D Morphable Face Model (FLAME) [43], which is decomposed

into facial expression coefficients $\alpha \in \mathbb{R}^{53}$ and pose coefficients $\beta \in \mathbb{R}^3$. Thus, each 3D head animation $f^S \in \mathbb{R}^{53+3}$ is formulated as:

$$f^S = \{\alpha, \beta\}. \quad (1)$$

For the previous mono audio-driven methods, the system output $\hat{f}_{1:t}^S$ can be formulated as:

$$\hat{f}^S = P_\theta(\mathcal{A}^S), \quad (2)$$

where $P_\theta(\cdot)$ indicates the generation model with learnable parameters θ . In this paper, considering a hybrid song as input, we propose a Partner-aware Chorus generation framework that addresses the challenge of generating interaction-driven singing animations for choral performances. Our model can be described as:

$$\begin{aligned} \hat{f}^{S1} &= P_\theta(\mathcal{A}^H = \mathcal{A}^{S1}, \mathcal{A}^P = \mathcal{A}^{S2}, \mathcal{A}^B), \\ \hat{f}^{S2} &= P_\theta(\mathcal{A}^H = \mathcal{A}^{S2}, \mathcal{A}^P = \mathcal{A}^{S1}, \mathcal{A}^B), \end{aligned} \quad (3)$$

where \mathcal{A}^H indicates the host vocal, \mathcal{A}^P indicates the partner vocal. If three or more singers are involved in the song, we mixture all the partner vocals except the host vocal into one track, that is $\mathcal{A}^P = \mathcal{A}^S - \mathcal{A}^H$.

4.2. Partner-aware Chorus Generation Framework

In this section, we propose a Partner-aware Chorus Generation Framework, PaChorus. As shown in Fig. 2, the framework is designed with two branches and end-to-end training. The first branch Partner-aware Motion Prior Learning (upper left of Fig. 2), takes the partners’ vocal and background music as inputs and predicts the pose coefficients β only. At the same time, the second branch Multi-Singers Animator (bottom left of Fig. 2)), extracts features from the host vocal and aims to predict all the FLAME coefficients f^S . Specifically, the two-branch framework is a reformulation of Eq. (3):

$$\begin{aligned} \Delta \hat{\beta}^S &= P_{\theta_1}(\mathcal{A}^P, \mathcal{A}^B), \\ \hat{f}^S &= P_{\theta_2}(\mathcal{A}^H), \end{aligned} \quad (4)$$

where θ_1 and θ_2 are the learnable parameters in the first and second branches, respectively. The full framework design is based on the insight that while the host vocal should control all the animation coefficients, the partner vocals as well as the background music should only provide emotion priors and affect the host’s head movement. In the following, we introduce the Host Animator and the Partner-aware Motion Prior in details.

4.3. Host Animator

To generate lip and facial animation fulfilling the condition of host’s vocal, Host Animator utilizes the host audio cues to generate synchronized facial movements for solo singing performances. Following established practices in

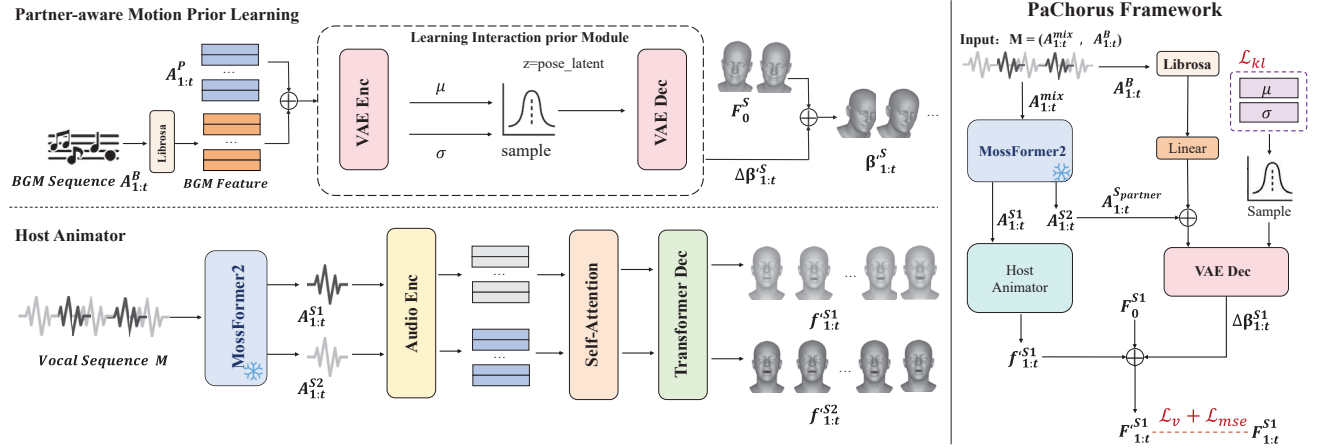


Figure 2. Overview of PaChorus framework. PaChorus includes two branches: Partner-aware Motion Prior Learning and Host Animator. The prior learning branch is designed for pose animation generation based on the modeling of inter-singers interaction and music, and the host animator branch is introduced to generate faithful lip and facial animation conditioned on the host’s vocal.

previous research, we use Wav2Vec [44] as the audio encoder. The separated audio is first processed through a pre-trained Wav2Vec2.0 model to extract content-related features, which are then mapped to lip movements. This allows us to capture detailed voice characteristics and their temporal dynamics. The extracted audio features are further encoded via a learned transformation network, mapping them into the desired latent space suitable for facial animation.

To model the lip movements, we employ a lip mask extracted from the FLAME model, which focuses on the regions of the face responsible for lip movements. The lip motion features are passed through a linear layer, and temporal alignment is ensured through linear interpolation. This allows the model to map varying input frame rates into a consistent output frame rate suitable for generating smooth lip movements that match the corresponding audio.

The overall structure of the decoder is based on a self-attention layer and transformer, using the previously encoded audio features to predict a sequence of coefficients that represent both head and facial movements. These coefficients are subsequently converted into vertices using the FLAME model to represent the 3D facial structure. By applying a residual transformation, we ensure that the output vertices can incorporate personalized variations while maintaining smooth and coherent facial animations.

4.4. Learning Partner-aware Motion Prior

Based on our observations of choral performances within ChorusHead dataset, we found that when one singer is actively singing, the other tends to respond with subtle facial movements, such as turning toward their partner and sharing a similar emotional expression. This suggests that human interactions in a group are not solely dictated by the

individual performances, but also by how performers react to each other, creating a sense of synchrony and interaction. Such reactions are often influenced by both the partner’s singing and the background music.

To capture these behaviors, we propose a partner-aware motion prior learning that models head pose generation based on these interpersonal dynamics. Due to the head movement diversity, we model the motion prior in probability. Specifically, we adopt a VAE-based [45] framework to learn a partner-aware interactive latent space. A conditional variational autoencoder (CVAE) model [46] is adopted to construct our partner-aware interaction module, which is built on transformer [47] layers and conditioned on background music and partner’s audio embeddings. During training, the encoder of the CVAE maps the head poses into a latent Gaussian space, producing a distribution $q(z|x)$ with a mean μ and variance σ^2 . The latent variable z is sampled from this distribution, defined as:

$$z \sim \mathcal{N}(\mu, \sigma^2). \quad (5)$$

In the decoder, the network learns to generate head poses from the sampled latent distribution. Crucially, instead of generating absolute poses directly, our model predicts the residual pose relative to the initial pose F_0^S from the first frame. This residual-based prediction ensures that head motion remains smooth, continuous, and stable over longer sequences during inference:

$$\hat{\beta}_{1:t} = F_0^S + \Delta\beta_{1:t}. \quad (6)$$

Furthermore, we incorporate the features from singing partner’s audio \mathcal{A}^P and background music \mathcal{A}^B into the CVAE conditioning, enabling the module to adapt to both rhythmic patterns and the partner’s individual voice.

4.5. Optimization

The optimization process involves the mean square error loss \mathcal{L}_{mse} to ensure animation generation accuracy and KL-divergence loss \mathcal{L}_{kl} to align the latent distribution in VAE:

$$\begin{aligned}\mathcal{L}_{mse} &= \frac{1}{t} \sum_{i=1}^t \|\hat{f}_i - f_i^{gt}\|^2, \\ \mathcal{L}_{kl} &= \sum(\sigma^2 + \mu^2 - \log(\sigma^2) - 1).\end{aligned}\quad (7)$$

Also, we adopt the velocity loss \mathcal{L}_v same to the previous methods [4, 27], to improve temporal consistency. The overall loss function can be formulated as:

$$\mathcal{L}_{all} = w_1 \mathcal{L}_{mse} + w_2 \mathcal{L}_{kl} + w_3 \mathcal{L}_v, \quad (8)$$

where $w_1 = w_3 = 1$, $w_2 = 0.005$ is set empirically. During the training, the parameters in MossFormer2 and TCN of wav2vec is frozen, and all the other model’s parameters are trained together in an end-to-end manner.

5. Experiment

5.1. Implementation details

We implemented the framework *PaChorus* in PyTorch [48], and all the training process are executed on a NVIDIA RTX3090 GPU. We trained our model for 100 epochs using Adam [49] optimizer, with learning rate setting to 0.0001. We conducted experiments on the ChorusHead dataset, which was divided into training, validation, and testing sets with splits of 60%, 10%, and 30%, respectively. Each sample was segmented into audio clips and corresponding action sequences for singer1 and singer2 using a 5-second sliding window with a 0.5-second stride. To accelerate data loading, we stored the processed data in LMDB format. For each method compared in the experiments, we follow the official setting to train and test their model with the least of modification to adapt to our dataset by generating each singer’s motion based on separate vocal audio condition.

5.2. Experimental Setup and Evaluation Metrics

To assess the effectiveness of our multi-person singing animation generation, we evaluate both lip synchronization and overall facial realism as mainstream audio-driven facial animation methods [1–4, 29].

Lip synchronization. We assess the alignment of generated facial movements by comparing vertex positions between generated animations and ground truth. We use multiple metrics to evaluate our experiments:

- Lip Vertex Error (LVE) calculates the average maximum L2 distance across all lip vertices per frame.
- Face Vertex Error (FVE) calculates the average maximum L2 distance among face region vertices.
- Face Dynamic Deviation (FDD) calculate the upper facial dynamic deviation of upper face region.

Table 2. Evaluation results of different methods on the ChorusHead, showing performance on LVE, FVE, and FDD. Lower values indicate better performance.

	Method	LVE ↓	FVE ↓	FDD ↓
Pretrained	FaceFormer [1]	2.0489	1.4862	1.4865
	CodeTalker [2]	2.2691	1.9294	2.7021
	Imitator [4]	2.2754	1.7211	1.8558
	SelfTalk [3]	1.7956	1.6522	2.7456
Finetuned	FaceFormer [1]	1.5883	1.4657	2.3542
	CodeTalker [2]	1.7676	1.8626	2.4829
	Imitator [4]	1.8125	1.4219	2.3267
	SelfTalk [3]	1.6687	1.5624	1.9458
	Diffspeaker [29]	1.5833	1.5672	2.2121
	PaChorus (Ours)	1.1922	1.3083	1.2314

Head movement. Most previous work has overlooked the statistical evaluation of head movement metrics, which are particularly crucial for generating realistic and expressive singing face animations. The metric we applied:

- Pose Parameters Error (PPE) compute the L2 distance between the predicted and ground truth global head rotation parameters.

5.3. Comparison to Existing Methods

Currently, there are no existing methods tailored for song-driven animations involving multiple singers. To ensure experimental fairness, we extend the single-person animation approach to multi-person datasets. All methods are finetuned on our ChorusHead dataset, enabling a fair comparison with the results. We then benchmark our proposed model against several 3D animation generation methods, evaluating its performance across multiple metrics to showcase its strengths and highlight differences from other approaches. Notably, most existing methods are designed specifically for mesh sequences and predict vertex offsets directly, which limits their accuracy in modeling head rotations. To ensure fairness in comparison, we evaluated all lip-related metrics with a fixed head posture.

As the quantitative results are shown in Tab. 2, *PaChorus* outperforms the mono audio-driven methods. The pre-trained models on talking datasets exhibit the lowest performance across multiple metrics, including LVE, FVE, and FDD. This suggests a domain gap between talking and singing, indicating that methods tailored for speech cannot be directly applied to singing contexts. Additionally, the results also show that our *PaChorus* achieves state-of-the-art performance across multiple metrics.

5.4. Ablation study

To evaluate the effectiveness of our proposed model components for multi-person song-driven head animation based on hybrid singing voice, we conducted an ablation study. Each

Table 3. Comparison of model variations across different metrics.

Methods	LVE ↓	FVE ↓	FDD ↓	PPE ↓
w/o BGM	1.2519	1.3395	1.3985	0.1295
w/o \mathcal{L}_{kl}	1.2589	1.3241	1.3214	0.1461
w/o prior	1.2409	1.3419	1.3421	0.1246
PaChorus (Ours)	1.1922	1.3083	1.2314	0.0948

ablated variant of our model omits a specific component, and we analyze how this affects each metric to understand the contribution of each part to overall model performance. **Effectiveness of BGM.** As the result in Tab. 3, Background music serves as a rhythmic guide for the generated animations. Compare “w/o BGM” with “Ours”, the LVE of “w/o BGM” is slightly increased to original. Notably, background music contains rhythm and other related information, which influences the speed and timing of head movements. Additionally, the increase in PPE implies that the head pose consistency is also impacted due to the absence of rhythmic cues, leading to less expressive head movements. **Effectiveness of \mathcal{L}_{kl} .** The KL divergence loss (\mathcal{L}_{kl}) within the CVAE framework plays a crucial role in regularizing the latent space, ensuring that the distribution of latent variables aligns with a prior distribution. As indicated in Tab. 3, when the \mathcal{L}_{kl} is omitted, there is a noticeable increase in the performance metrics, particularly in PPE. This suggests that without the regularization effect of KL divergence, the model’s ability to maintain proper alignment is compromised. Specifically, the accuracy of head pose generation significantly decreases, leading to negative guidance on head movements.

Effectiveness of prior. The prior component within the CVAE framework captures cues from the background music’s rhythm and the partner’s speech features, enabling more expressive and synchronized animations. Omitting the prior results in increased LVE and FVE as shown in Tab. 3, indicating a loss in both lip and facial synchronization. Additionally, PPE increases, suggesting that the absence of the prior diminishes head pose accuracy.

5.5. Multi-singers Animation Generation

As shown in Fig. 3, we demonstrate our model’s ability to generate multi-singer animations with synchronized lip movements for each individual. Due to current limitations in multi-speaker audio separation models, which are not yet fully reliable for extracting voices from group singing, we simulate separated vocal tracks by using re-recorded covers of the same song by different singers. These pre-separated tracks serve as input for our inference pipeline, allowing each track to drive synchronized lip movements through the Host Animator module. The distinct vocal characteristics of each singer naturally lead to subtle variations in lip shapes for each syllable, enhancing the animation’s vibrancy and

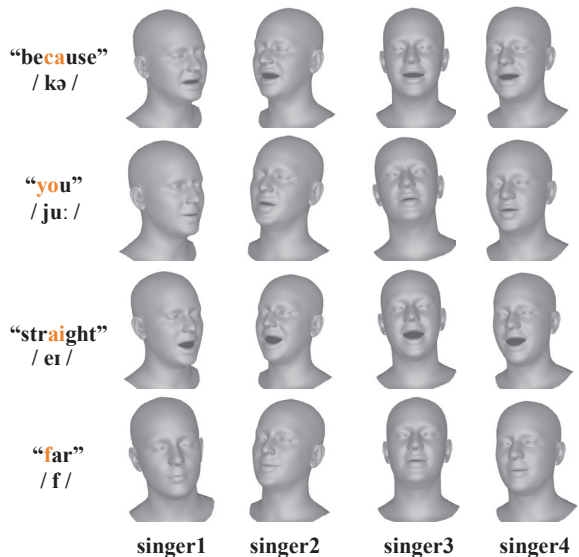


Figure 3. Multi-singer choral performance of song *because of you*.

Table 4. User study results from Amazon Mechanical Turk.

Method	PaChorus Win Rate (%)
FaceFormer	80.0
CodeTalker	86.7
Imitator	93.3
SelfTalk	83.3
Diffspeaker	78.6

realism. By incorporating background music and partners’ vocal features within the Motion Prior module, we achieve partner-aware interactive head poses, capturing responsive, musically aligned head dynamics between singers.

5.6. Qualitative Evaluation

To further assess the perceptual quality of our method, we conducted a user study via Amazon Mechanical Turk. 30 participants assessed animations rendered in grayscale to minimize visual distractions and direct attention to motion dynamics and audio-visual synchronization. From our test dataset, we randomly selected 50 samples and generated animations using our method and other finetuned baseline approaches. Evaluators were instructed to select the better animations based on two criteria: 1) lip synchronization accuracy and 2) head motion naturalness during interaction. As demonstrated in Tab. 4 and Fig. 4, our method demonstrates statistically significant superiority over competing approaches.

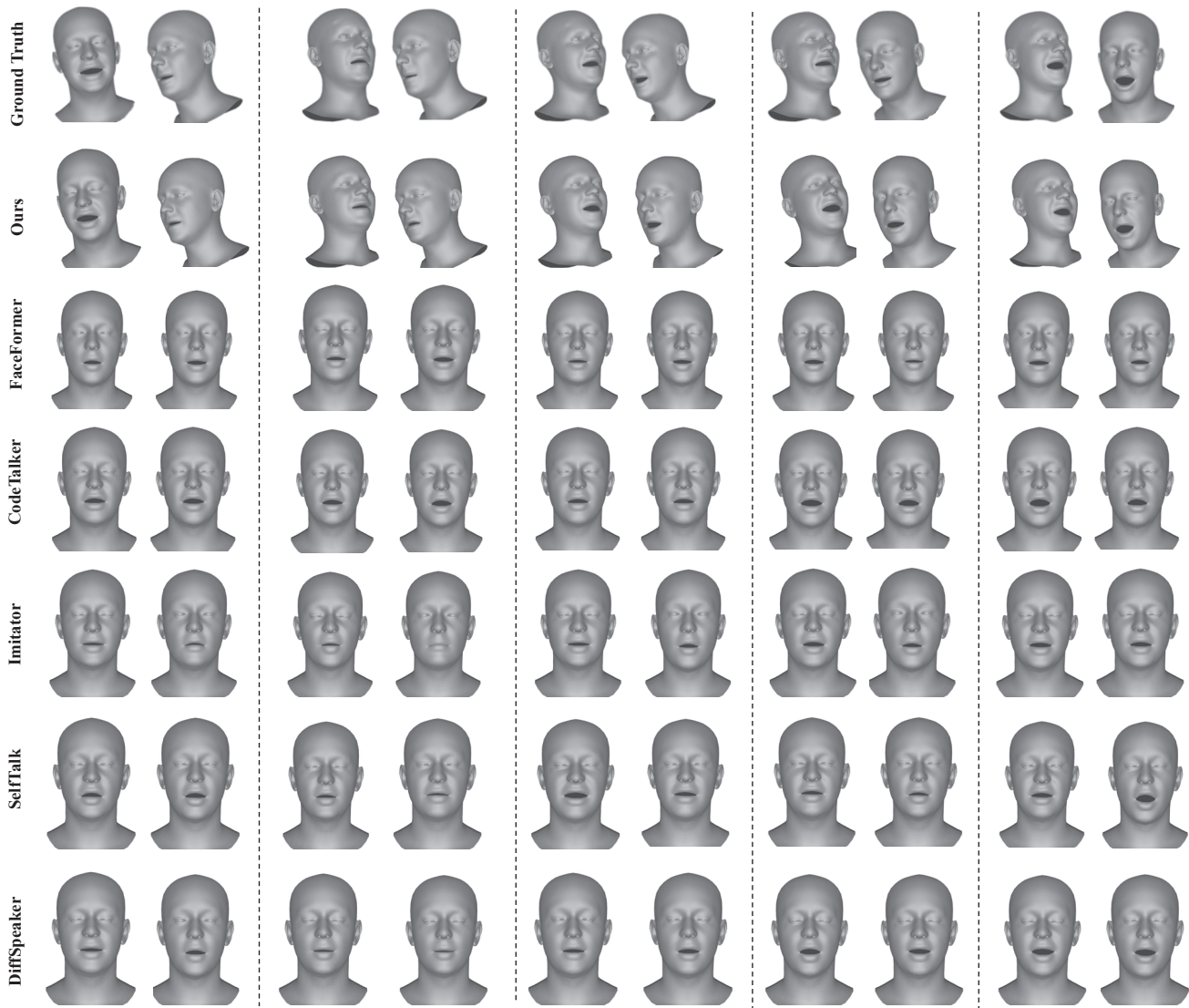


Figure 4. Visualization comparison between existing audio-driven 3D animation methods.

6. Conclusion

In this work, we introduced a novel framework, PaChorus, designed for multi-singer facial animation driven by a choral song, focusing on generating expressive and synchronized singing animations. Unlike traditional single-speaker animation models, our framework addresses the unique challenges of multi-person interactions and generate more vivid animation. By integrating components that model both mutual interactions between singers and rhythmic synchronization with background music, PaChorus produces realistic animations where singers’ expressions, head movements, and lip synchronization align seamlessly with the audio input. Our quantitative and qualitative

evaluations demonstrate the effectiveness of the proposed framework, showing improved performance across various metrics. Extend application also shows our ability to generate multi-singers animation.

Limitation. The absence of an effective audio decomposition model for 3 or more singers limit the data collection process on a larger scope. We will keep seeking a more powerful tool, or try to construct a new pipeline that do not rely on audio-decomposition in the future.

Broader Impact. We set a new benchmark for multi-singer facial animation, paving the way for more immersive and interactive virtual performances. We will release the dataset once published, and it may inspire further exploration of mixed audio-driven methods in the community.

Acknowledgements

The work is supported by National Natural Science Foundation of China (No.62302170), Guangdong Basic and Applied Basic Research Foundation (No.2024A1515010187), Guangdong Provincial Natural Science Foundation for Outstanding Youth Team Project (No.2024B1515040010), China National Key R&D Program (No.2023YFE0202700, 2024YFB4709200), Key-Area Research and Development Program of Guangzhou City (No.2023B01J0022), NSFC Key Project (No.U23A20391)

References

- [1] Yingruo Fan, Zhaojiang Lin, Jun Saito, Wenping Wang, and Taku Komura. Faceformer: Speech-driven 3d facial animation with transformers. In *CVPR*, pages 18770–18780, 2022. [2](#), [3](#), [6](#)
- [2] 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 *CVPR*, pages 12780–12790, 2023. [3](#), [6](#)
- [3] 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 *ACM MM*, pages 5292–5301, 2023. [6](#)
- [4] Balamurugan Thambiraja, Ikhsanul Habibie, Sadegh Aliakbarian, Darren Cosker, Christian Theobald, and Justus Thies. Imitator: Personalized speech-driven 3d facial animation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20621–20631, 2023. [2](#), [3](#), [6](#)
- [5] Yifang Pan, Chris Landreth, Eugene Fiume, and Karan Singh. Vocal: Vowel and consonant layering for expressive animator-centric singing animation. In *SIGGRAPH Asia*, New York, NY, USA, 2022. Association for Computing Machinery. [2](#), [3](#)
- [6] Shohei Iwase, Takuya Kato, Shugo Yamaguchi, Tsuchiya Yukitaka, and Shigeo Morishima. Song2face: Synthesizing singing facial animation from audio. In *SIGGRAPH Asia*, New York, NY, USA, 2020. Association for Computing Machinery. [3](#), [4](#)
- [7] Pengfei Liu, Wenjin Deng, Hengda Li, Jintai Wang, Yinglin Zheng, Yiwei Ding, Xiaohu Guo, and Ming Zeng. Music-face: Music-driven expressive singing face synthesis. *Computational Visual Media*, 10(1):119–136, 2024. [3](#), [4](#)
- [8] Sijing Wu, Yunhao Li, Weitian Zhang, Jun Jia, Yucheng Zhu, Yichao Yan, Guangtao Zhai, and Xiaokang Yang. Singing-head: A large-scale 4d dataset for singing head animation. *arXiv preprint arXiv:2312.04369*, 2023. [2](#), [3](#), [4](#)
- [9] Ariel Ephrat, Inbar Mosseri, Oran Lang, Tali Dekel, Kevin Wilson, Avinatan Hassidim, William T. Freeman, and Michael Rubinstein. Looking to listen at the cocktail party: a speaker-independent audio-visual model for speech separation. *ACM TOG*, 37(4):1–11, July 2018. [2](#), [3](#)
- [10] Ruohan Gao and Kristen Grauman. Visualvoice: Audio-visual speech separation with cross-modal consistency. In *CVPR*, pages 15490–15500. IEEE, 2021.
- [11] Neil Zeghidour and David Grangier. Wavesplit: End-to-end speech separation by speaker clustering. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:2840–2849, 2021. [3](#)
- [12] Juan F Montesinos, Venkatesh S Kadandale, and Gloria Haro. A cappella: Audio-visual singing voice separation. *arXiv preprint arXiv:2104.09946*, 2021. [2](#)
- [13] Sicheng Yang, Zhiyong Wu, Minglei Li, Zhensong Zhang, Lei Hao, Weihong Bao, Ming Cheng, and Long Xiao. Diffusestylegesture: Stylized audio-driven co-speech gesture generation with diffusion models. *arXiv preprint arXiv:2305.04919*, 2023. [2](#)
- [14] Hao Meng, Sheng Jin, Wentao Liu, Chen Qian, Mengxiang Lin, Wanli Ouyang, and Ping Luo. 3d interacting hand pose estimation by hand de-occlusion and removal. In *European Conference on Computer Vision*, pages 380–397. Springer, 2022.
- [15] Kiran Chhatre, Nikos Athanasiou, Giorgio Becherini, Christopher Peters, Michael J Black, Timo Bolkart, et al. Emotional speech-driven 3d body animation via disentangled latent diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1942–1953, 2024.
- [16] Hongze Yao, Yingting Xu, Weitao WU, Huabin He, Wen Ren, and Zhiming Cai. Audio2diffugesture: Generating a diverse co-speech gesture based on a diffusion model. *Electronic Research Archive*, 32(9), 2024. [2](#)
- [17] Xinya Ji, Hang Zhou, Kaisiyuan Wang, Qianyi Wu, Wayne Wu, Feng Xu, and Xun Cao. Eamm: One-shot emotional talking face via audio-based emotion-aware motion model. In *ACM Trans. Graph.*, pages 1–10, 2022. [2](#)
- [18] Xinya Ji, Hang Zhou, Kaisiyuan Wang, Wayne Wu, Chen Change Loy, Xun Cao, and Feng Xu. Audio-driven emotional video portraits. In *CVPR*, pages 14080–14089, 2021. [2](#)
- [19] Haozhe Wu, Jia Jia, Junliang Xing, Hongwei Xu, Xiangyuan Wang, and Jelo Wang. Mmface4d: A large-scale multi-modal 4d face dataset for audio-driven 3d face animation. *arXiv preprint arXiv:2303.09797*, 2023. [2](#)
- [20] Pif Edwards, Chris Landreth, Eugene Fiume, and Karan Singh. Jali: an animator-centric viseme model for expressive lip synchronization. *ACM TOG*, 35(4), July 2016. [2](#)
- [21] Yang Zhou, Zhan Xu, Chris Landreth, Evangelos Kalogerakis, Subhransu Maji, and Karan Singh. Visemenet: Audio-driven animator-centric speech animation. *ACM TOG*, 37(4):1–10, 2018.
- [22] Sarah L. Taylor, Moshe Mahler, Barry-John Theobald, and Iain Matthews. Dynamic units of visual speech. In *Proceedings of the ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, page 275–284, Goslar, DEU, 2012. Eurographics Association. [2](#)
- [23] Yudong Guo, Keyu Chen, Sen Liang, Yong-Jin Liu, Hujun Bao, and Juyong Zhang. Ad-nerf: Audio driven neural radiance fields for talking head synthesis. In *ICCV*, pages 5784–5794, 2021. [3](#)
- [24] Tero Karras, Timo Aila, Samuli Laine, Antti Herva, and Jaakko Lehtinen. Audio-driven facial animation by joint end-

- to-end learning of pose and emotion. *ACM TOG*, 36(4):1–12, 2017.
- [25] Avisek Lahiri, Vivek Kwatra, Christian Frueh, John Lewis, and Chris Bregler. Lipsync3d: Data-efficient learning of personalized 3d talking faces from video using pose and lighting normalization. In *CVPR*, pages 2755–2764, 2021. 3
- [26] Shunyu Yao, RuiZhe Zhong, Yichao Yan, Guangtao Zhai, and Xiaokang Yang. Dfa-nerf: Personalized talking head generation via disentangled face attributes neural rendering. *arXiv preprint arXiv:2201.00791*, 2022. 3
- [27] Daniel Cudeiro, Timo Bolkart, Cassidy Laidlaw, Anurag Ranjan, and Michael J Black. Capture, learning, and synthesis of 3d speaking styles. In *CVPR*, pages 10101–10111, 2019. 3, 4, 6
- [28] Stefan Stan, Kazi Injamamul Haque, and Zerrin Yumak. Facediffuser: Speech-driven 3d facial animation synthesis using diffusion. In *Proceedings of the 16th ACM SIGGRAPH Conference on Motion, Interaction and Games*, pages 1–11, 2023. 3
- [29] Zhiyuan Ma, Xiangyu Zhu, Guojun Qi, Chen Qian, Zhaoxiang Zhang, and Zhen Lei. Diffspeaker: Speech-driven 3d facial animation with diffusion transformer. *arXiv preprint arXiv:2402.05712*, 2024. 6
- [30] 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 TOG*, 43(4):1–9, 2024. 3
- [31] 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 *ICCV*, pages 20687–20697, 2023. 3, 4
- [32] Radek Daněček, Kiran Chhatre, Shashank Tripathi, Yandong Wen, Michael Black, and Timo Bolkart. Emotional speech-driven animation with content-emotion disentanglement. In *SIGGRAPH Asia*, page 1–13. ACM, December 2023. 3
- [33] Yicheng Zhong, Huawei Wei, Peiji Yang, and Zhisheng Wang. Expclip: Bridging text and facial expressions via semantic alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 7614–7622, 2024. 3
- [34] Robin Scheibler, Youna Ji, Soo-Whan Chung, Jaek Byun, Soyeon Choe, and Min-Seok Choi. Diffusion-based generative speech source separation. In *ICASSP*, pages 1–5. IEEE, 2023. 3
- [35] Shengkui Zhao, Yukun Ma, Chongjia Ni, Chong Zhang, Hao Wang, Trung Hieu Nguyen, Kun Zhou, Jia Qi Yip, Dianwen Ng, and Bin Ma. Mossformer2: Combining transformer and rnn-free recurrent network for enhanced time-domain monaural speech separation. In *ICASSP*, pages 10356–10360. IEEE, 2024. 3, 4
- [36] Romain Hennequin, Anis Khelif, Felix Voituret, and Manuel Moussallam. Spleeter: a fast and efficient music source separation tool with pre-trained models. *Journal of Open Source Software*, 5(50):2154, 2020. Deezer Research. 3, 4
- [37] Bo Jiang, Shaoyu Chen, Qing Xu, Bencheng Liao, Jiajie Chen, Helong Zhou, Qian Zhang, Wenyu Liu, Chang Huang, and Xinggong Wang. Vad: Vectorized scene representation for efficient autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8340–8350, 2023. 4
- [38] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, October 2016. 4
- [39] Radek Daněček, Michael J Black, and Timo Bolkart. Emoca: Emotion driven monocular face capture and animation. In *CVPR*, pages 20311–20322, 2022. 4
- [40] Kaisiyuan Wang, Qianyi Wu, Linsen Song, Zhuoqian Yang, Wayne Wu, Chen Qian, Ran He, Yu Qiao, and Chen Change Loy. Mead: A large-scale audio-visual dataset for emotional talking-face generation. In *ECCV*, 2020. 4
- [41] Zhimeng Zhang, Lincheng Li, Yu Ding, and Changjie Fan. Flow-guided one-shot talking face generation with a high-resolution audio-visual dataset. In *CVPR*, pages 3661–3670, 2021. 4
- [42] Steven R Livingstone and Frank A Russo. The ryer-son 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):e0196391, 2018. 4
- [43] Tianye Li, Timo Bolkart, Michael J Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4d scans. *ACM TOG*, 36(6):194–1, 2017. 4
- [44] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *NeurIPS*, 33:12449–12460, 2020. 5
- [45] Diederik P Kingma. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 5
- [46] Stepan Yu Orevkov. Automorphism group of the commutator subgroup of the braid group. In *Annales de la Faculte des sciences de Toulouse: Mathematiques*, volume 26, pages 1137–1161, 2017. 5
- [47] A Vaswani. Attention is all you need. *NeurIPS*, 2017. 5
- [48] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *NeurIPS*, 32, 2019. 6
- [49] Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6