MagicTalk: *Implicit* and *Explicit* Correlation Learning for Diffusion-based Emotional Talking Face Generation

Chenxu Zhang*
ByteDance Inc.
chenxuzhang@bytedance.com

Chao Wang*
ByteDance Inc.
chaowang15@gmail.com

Jianfeng Zhang
ByteDance Inc.
jianfengzhang@bytedance.com

Hongyi Xu
ByteDance Inc.
hongyixu@bytedance.com

Guoxian Song
ByteDance Inc.
guoxiansong@bytedance.com

You Xie
ByteDance Inc.
you.xie@bytedance.com

Linjie Luo
ByteDance Inc.
linjie.luo@bytedance.com

Yapeng Tian
University of Texas at Dallas
yapeng.tian@utdallas.edu

Jiashi Feng
ByteDance Inc.
jshfeng@bytedance.com

Xiaohu Guo
University of Texas at Dallas
xguo@utdallas.edu

Abstract

Generating emotional talking faces from a single portrait image remains a significant challenge. The simultaneous achievement of expressive emotional talking and accurate lip-sync is particularly difficult, as expressiveness is often compromised for lip-sync accuracy. Prevailing generative works usually struggle to juggle subtle variations of emotional expression and lip-synchronized talking generation. To address these challenges, we argue to model the implicit and explicit correlations between audio and emotional talking faces with a unified framework. As human emotional expressions usually present subtle and implicit relations with speech audio, we propose incorporating audio and emotional style embeddings into the diffusion-based generation process, indicating the realistic generation while concentrating on emotional expressions. We then propose lip-based explicit correlation learning to construct a strong mapping of audio and lip motions, assuring the lip-audio synchronizations. Besides, we deploy a video-to-video rendering module to transfer the expressions and lip motions from our proxy 3D avatar to an arbitrary portrait. Both quantitatively and qualitatively, MagicTalk outperforms state-of-the-art methods in terms of expressiveness, lip-sync and perceptual quality.

Keywords: ***Emotional talking face, Diffusion model,***

*Equal contribution

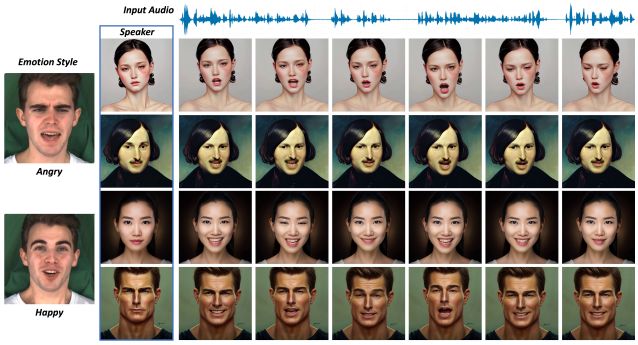

Figure 1: **MagicTalk** takes as input a driving audio sequence, a given portrait image, and an example of emotion style (a clip of an emotional talking face), and generates a photorealistic, lip-synchronized talking face video that features high-quality emotional expressions. The results include both real human images and images generated by AIGC. Please refer to our Project Page for more results.

Single image, Implicit and explicit correlation learning.

1. Introduction

The field of talking face generation has seen significant advancements in recent years, becoming a key area of research with a wide range of applications, including video conferencing, virtual assistants, and entertainment, among others. Recently, researchers have commenced incorporating emotion-conditioned facial expressions into talking face

generation [19, 18, 14], leveraging emotional annotations in talking video datasets. However, none of these methods have yet succeeded in generating expressive and lifelike expressions in their talking faces.

Several major challenges exist in current emotional talking face generation methods. Firstly, it is difficult to achieve *expressive emotion* and *accurate lip-sync* simultaneously. Emotional expressions in datasets like MEAD [51] show significant exaggeration in the movements of eyebrows, eye-blinking, and mouth shapes. Nonetheless, the sentences and audio content used in these datasets lack sufficient length to effectively train a precise lip-sync model. To address this issue, SPACE [14] employed the supplementary non-emotional dataset, namely VoxCeleb2 [6], alongside emotional datasets [27, 51], to train their model. However, integrating non-emotional and emotional datasets results in synthesized emotional expressions that may lack the desired level of expressiveness and dynamism. EAMM [18] tackles this problem by integrating two modules: one dedicated to learning non-emotional audio-driven face synthesis and another focused on capturing emotional displacements of expressions. To prevent emotional displacements from distorting lip-sync, it augmented the training data by obscuring the mouth region of the speakers. Unfortunately, employing mouth-covering data augmentation compromises the expressiveness of mouth shapes during emotional speech.

Secondly, modeling the subtleties and variations of emotional expressions is challenging. Emotional expressions involve the activation of numerous facial muscles and exhibit significant diversity throughout a speech. Existing methods [19, 18, 28, 14] typically utilize LSTM or CNN networks as generators to transform audio into facial representations. While these models are adequate for capturing the movements of the mouth and lips during regular speech, they face challenges when it comes to faithfully portraying the nuances and variations of emotional expressions. Consequently, their generated emotional depictions often appear bland and artificial. Unlike lip-sync, which has a strict frame-by-frame alignment with audio, the relationship between emotional expressions and audio is implicit and represents a global state correlation. This requires overall realism and consistency rather than precise per-frame matching. For example, sad emotion is characterized by a slower speech rate and softer volume, accompanied by a downward gaze and small head movements. However, current models often fail to capture these holistic emotional cues.

To overcome these challenges, we introduce a collaborative implicit-explicit correlation learning framework called MagicTalk. This framework aims to generate realistic emotional expressions by leveraging diffusion-based learning of implicit information such as blinking and furrowing brows through implicit correlations with audio. Subsequently, we generate lip motion for lip sync by learning precise mouth movements through explicit correlations with audio. At its core is a carefully designed implicit-explicit correlation learning pipeline, which achieves both expressive emotion and precise lip-sync, as shown in Fig. 2. The first stage, *implicit correlation learning*, is tailored to capture the dynamic nature of emotional expressions. Specifically, we designed an emotion-conditioned diffusion model to transform input audio to the facial expressions of the ARKit model [25]. The second stage, *explicit correlation learning*, focuses on ensuring the precision of lip-sync in the generated talking faces. To enhance the synchronization of mouth movements with audio signals while preserving the richness of emotional expressions, we have developed a novel lip refinement network to re-optimize the parameters of the mouth based on audio signals and specific emotional styles. Unlike traditional face model [23] where mouth parameters are integrated with other facial parameters, using 3D ARKit model enables explicitly optimizing lip motion, ensuring that the intensity of other facial expressions remains unaffected. This design choice in our lip refinement network guarantees that the expressiveness of emotions is not compromised by lip-sync refinement, offering a more targeted and effective approach for emotion-rich facial animation.

The sequential implicit-explicit correlation learning process employed in MagicTalk effectively addresses the challenges mentioned earlier, allowing for the simultaneous achievement of expressive emotions and precise lip-sync in the generated talking faces. Our experimental results convincingly showcase its exceptional ability to model the intricacies and variations of emotional expressions from the input audio. This includes realistically capturing emotional movements in areas such as eyebrows, eye blinks, and beyond. Specifically, our diffusion model adeptly captures high-frequency facial details, while lip refinement further elevates the precision of mouth motion. The contributions of this paper can be summarized as follows:

- We introduce a collaborative implicit-explicit correlation learning to model the weak and strong relationships of audio inputs with emotional talking heads.
- We propose to learn the implicit correlation between audio and emotional expressions by gradually incorporating audio conditions with emotional style embeddings into the diffusion process of talking head generation.
- We present a lip-sync explicit correlation learning with refined mouth parameters optimization to capture acoustically aligned lip motion with expressive emotional talking faces.

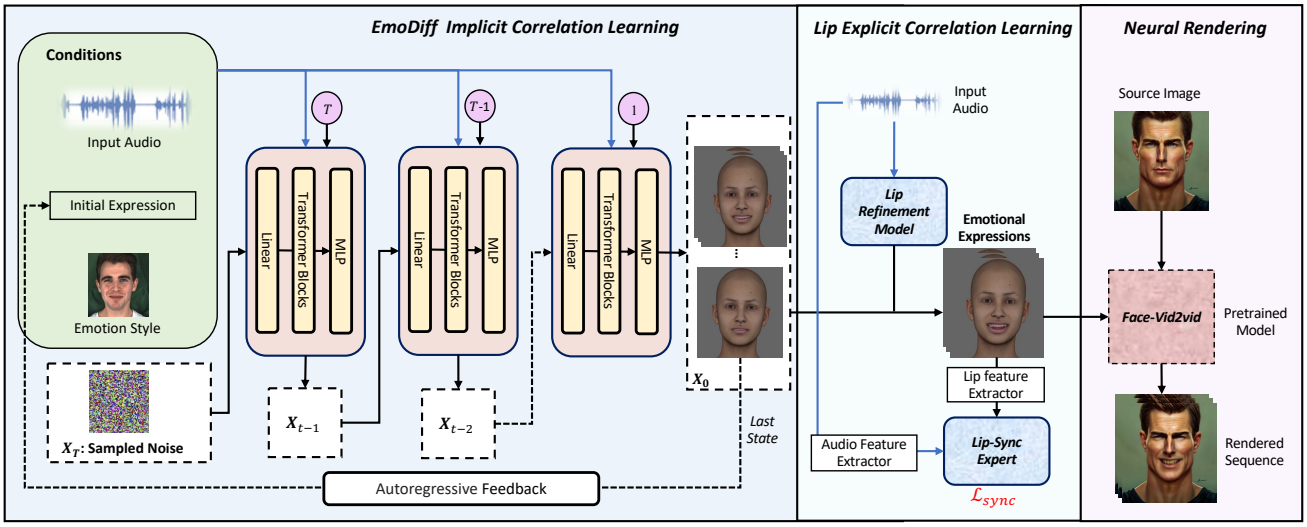

Figure 2: Pipeline of our MagicTalk framework. Starting with the input audio, initial state, and emotion style as conditions, we first employ EmoDiff for learning to denoise 3D expressions over time, utilizing a transformer-based architecture for sequence modeling. The initial state corresponds to the expression in the first frame, and the emotion style is defined by a randomly selected expression clip, independent of the input audio. Then, utilizing the conditioned audio and emotional expressions, the lip refinement model, in conjunction with the Lip-sync expert, enhances mouth movements without affecting the emotional intensity. This is followed by producing corresponding 3D rendering faces on a blendshape rig. Finally, we employ a fine-tuned Face-Vid2Vid model [52] to generate emotional talking videos.

2. Related Work

Audio-driven talking face. The task of generating a talking face driven by audio [59, 3, 50, 49, 39, 12, 8, 44, 45, 55, 24, 64, 61, 62] involves producing a realistic and cohesive video of a person's speaking face, utilizing audio input and, occasionally, an image or video of the speaker. Early efforts by Taylor et al. [44] focused on converting audio sequences into phoneme sequences to generate adaptable talking avatars for multiple languages. [40] curated a dataset of Obama videos and introduced an end-to-end framework to synthesize corresponding talking faces with arbitrary voices. [5] and [63] pioneered the generation of talking face videos, requiring only a single facial and audio sequence. ATVGnet [4] and [60] proposed two-stage talking face synthesis methods guided by landmarks. They initially generated landmarks from a single identity image and an audio sequence, which were then combined with the identity image for the second stage of talking face synthesis. [30] decomposed talking head synthesis into spatial and style components, demonstrating improved performance in few-shot novel view synthesis. Many methods [57, 37, 45, 42, 37, 22, 21, 56] used audio to regress parameters in 3D face models, resulting in more realistic synthesis. To enhance video quality, recent NeRF-based methods [13, 54] employed an audio-driven neural radiance fields (NeRF) model to synthesize high-quality talking-head videos from audio input, surpassing existing GAN-based methods. SadTalker [58] adeptly generates emotive speech content by mapping audio inputs to 3DMM motion coefficients, but challenges remain in achieving both realistic expression and accurate lip movement. On the other hand, [38, 34, 26, 33, 43] applied diffusion models to avoid challenges of GAN-based methods, such as training difficulties, mode collapse, and facial distortion. However, those methods require extra motion sequence of the target individual to guide the video generation and avoid unnatural-looking motions. Moreover, utilizing diffusion model as foundational framework, DiffTalk also encountered challenges in maintaining temporal coherence within the mouth region. On the other hand, the works discussed above still lack emotional information guidance, leading to monotonous expressions in generated talking faces.

Emotional audio-driven talking face. In recent research endeavors, there has been a growing focus on the development of emotionally expressive talking faces [53, 18, 41, 11, 48, 46] or emotional talking mesh [31]. ExprGAN [9] introduced an expression control module that enables the synthesis of faces with diverse expressions, allowing for continuous adjustment of expression intensity. [10] presented a neural network system conditioned on categorical emotions, providing direct and flexible control over visual emotion ex-

pression. MEAD [51] enhanced the vividness of talking faces by curating the MEAD dataset, offering a baseline for emotional talking face generation. [19] proposed a groundbreaking method for emotional control in video-based talking face generation, incorporating a component for distilling content-agnostic emotion features. Addressing the challenge of the timbre gap, [22] introduced a framework for talking-head synthesis that generates facial expressions and corresponding head animations from textual inputs. EAMM [18] and EMMN [41] both utilized 2D keypoints displacement to synthesize the final emotional video, which can degrade the quality of generation. [24] presented a method for generating expressive talking heads with meticulous control over mouth shape, head pose, and emotional expression. [36] proposed an optical flow-guided texture generation network capable of rendering emotional talking face animations from a single image, regardless of the initial neutral emotion. SPACE [14] introduced a method decomposing the task into facial landmark prediction and emotion conditioning, resulting in talking face videos with elevated resolution and fine-grained controllability. In our work, we employ a diffusion model to predict expression sequences, yielding more expressive outcomes.

3. Method

3.1. Preliminaries

Contrary to 2D landmark-based methods, which are susceptible to head pose variations and often face challenges in maintaining consistent facial shape representation [63, 18], 3D modeling techniques offer shape-invariant information, thereby facilitating more realistic renderings that align with the actual three-dimensional structure of human faces. Traditional 3D models, such as 3D Morphable Models (3DMM) or FLAME, predominantly utilize Principal Component Analysis (PCA) to encapsulate facial features. While these parameters provide control over general facial appearance, they fall short in isolating specific facial attributes, such as eye blinking or lip movements. Given our objective to enhance the mouth region while concurrently preserving the expressiveness of other facial features, we have elected to employ ARKit blend shapes. This technology distinctly separates mouth-related parameters from other facial elements, thus enabling targeted optimization. The ARKit facial model comprises 52 distinct parameters, each representing unique facial features. It utilizes blend shapes based on the Facial Action Coding System (FACS), allowing each facial expression to activate specific facial regions (e.g., mouth area, eyes, eyebrows) independently and in a manner consistent with human facial anatomy [25]. This approach offers precise control over and optimization of various facial attributes, rendering it particularly well-suited for our specialized optimization requirements.

Subsequently, we conduct a comprehensive analysis of ARKit parameters on each frame within the MEAD emotion dataset, thereby extracting corresponding parameters. This process facilitates the creation of an ARKit-specific facial dataset, meticulously tailored to align with the emotional nuances of the MEAD dataset. We develop an emotion dataset that features fully disentangled 3D facial parameters. Such a development significantly amplifies the practical utility and applicability of emotion-based datasets in the field. For the convenience of the community, we plan to release our dataset publicly in the future.

3.2. EmoDiff Implicit Correlation Learning

Our goal is to generate 3D emotional expressions from audio. However, this task presents significant challenges that require innovative solutions. Firstly, mapping audio to expressions is a one-to-many problem, making it difficult to obtain dynamic and realistic expressions. Secondly, generating a sequence of 3D face expression parameters involves numerous issues, such as continuity and diversity. To address these challenges, we propose to learn the implicit correlation between audio and emotional expressions by gradually incorporating audio conditions with emotional style embeddings into the diffusion process of talking head generation.

Forward diffusion and reverse process. We adopt denoising diffusion probabilistic models (DDPM) [16], where a real data sample $\boldsymbol{x_0} \sim q(\boldsymbol{x})$ undergoes a forward noising process, defined as:

$$q(\boldsymbol{x}_t|\boldsymbol{x_0}) = \mathcal{N}(\sqrt{\bar{\alpha}_t}\boldsymbol{x_0}, (1-\bar{\alpha}_t)\boldsymbol{I}), \tag{1}$$

where $\bar{\alpha}_t \in (0,1)$ follows a decreasing schedule, making $\boldsymbol{x}_T \approx \mathcal{N}(0, \boldsymbol{I})$.

The reverse process reconstructs $\boldsymbol{x_0}$ from noise $\boldsymbol{x_T} \sim \mathcal{N}(0, \boldsymbol{I})$ by modeling:

$$p_\theta(x_{t-1}|x_t, c) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t, c), \beta_t I), \tag{2}$$

where c represents conditioning inputs (audio, initial state, emotion style). Following [65], we set $\beta_t I$ as time-dependent constants.

Training objective. Since x_t is available as the input to the model, we predict the gaussian noise ϵ instead of μ at time step t:

$$\mu_\theta(x_t, t, c) = \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(x_t, t, c)\right), \tag{3}$$

where ϵ_θ is a function approximator intended to predict ϵ from x_t. We optimize θ with the following objective, which works better by ignoring the weighting term introduced in Ho et al. [16]:

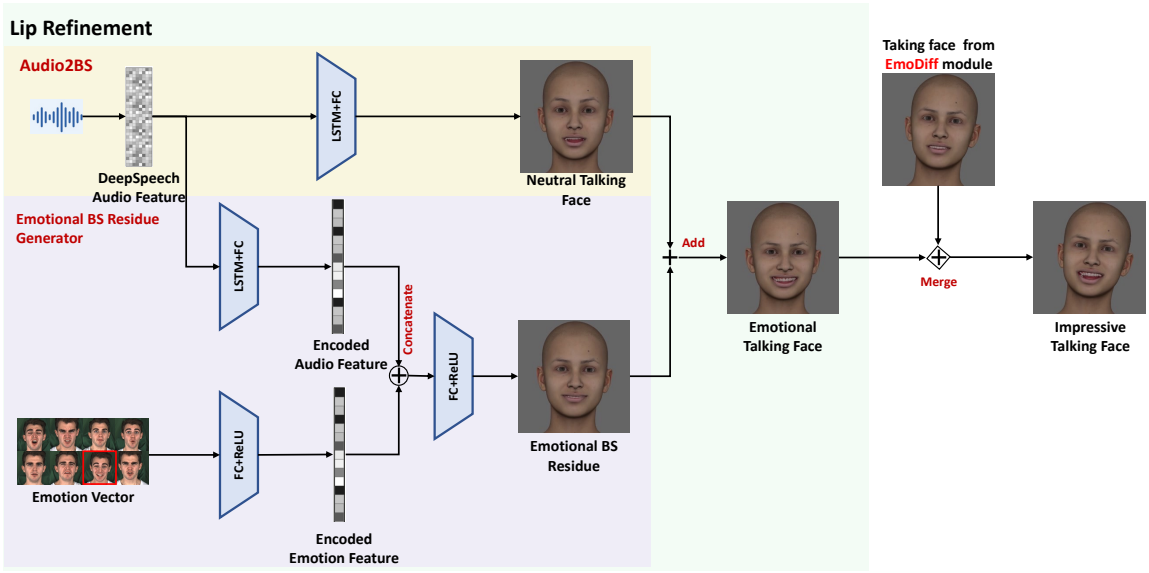

Figure 3: Pipeline of our Lip Refinement framework. Our Lip refinement network uses an LSTM structure to learn lip movements linked to audio. We start by creating a neutral talking face based on audio features, focusing on lip movements related to the content. We then add emotional elements by combining audio features with emotion features, producing an emotional Blendshape (BS) residue. This approach allows us to create a talking face with emotions that align closely with the audio. Finally, we enhance the face with head poses from the emodiff module and improve the lip-sync by replacing the emodiff's lip part with the one from our lip refinement network. The result is a more emotionally expressive talking face with better lip-syncing.

$$\begin{aligned}\mathcal{L}_{\text{simple}}(\theta) =& \mathbb{E}_{t,\boldsymbol{x_0},\boldsymbol{\epsilon}}\left[\|\boldsymbol{\epsilon}-\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t,t,\boldsymbol{c})\|_2^2\right] \\ =& \mathbb{E}_{t,\boldsymbol{x_0},\boldsymbol{\epsilon}}\left[\|\boldsymbol{\epsilon}-\boldsymbol{\epsilon}_\theta(\sqrt{\bar{\alpha}_t}\boldsymbol{x}_0+\sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon},t,\boldsymbol{c})\|_2^2\right],\end{aligned} \tag{4}$$

where t is uniformly sampled between 1 and T.

Classifier-free guidance. We train our conditional diffusion model by applying classifier-free guidance [17], which is widely used in recent diffusion-based works [47, 65]. Specifically, the condition $\boldsymbol{c}$ gets discarded periodically at random by setting $\boldsymbol{c} = \phi$. The guided inference is then expressed as the weighted sum of unconditionally and conditionally generated samples:

$$\hat{\boldsymbol{\epsilon}}_\theta = (w+1)\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t,t,\boldsymbol{c}) - w(\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t,t,\phi), \tag{5}$$

where w is the scale parameter to trade off unconditionally and conditionally generated samples. The classifier-free guidance can achieve a good balance between quality and diversity.

Diffusion model. For the selection of the network architecture, we primarily considered two issues: 1) how to incorporate different modality data, such as audio and 3D expressions, and 2) how to generate the N-frame 3D face expression sequence different from images. As shown in Fig. 2, we choose to use the transformer structure, which fuses the representation of different modalities and captures the long-term dependency with a cross-attention mechanism following [65]. Please refer to the supplementary materials for detailed information on the network architecture.

Time-aware positional embedding. Generating facial expressions with temporal continuity requires consideration of two issues. Firstly, our training dataset only includes T-frame input data. However, during testing, we may require longer testing audio, so we need to consider the continuity between generated sequences. To address this issue, we use the first frame of the generated sequence as the input feature condition and add it to the network, which constrains the initial state of the generated sequence.

Additionally, we need to guide the network to capture the style of emotional expression. To achieve this, we use three frames of expression as a representation of emotion style. To prevent the style from representing audio information, we randomly sample three frames of expression during training. To incorporate the initial state condition and style information into the network, we use a time position-aware approach. Specifically, we first create a matrix with the same length as the audio frames and set the first column of the matrix to the initial state value, a 50-dimensional vec-

Algorithm 1 Long-term Dynamic Sampling

Trained diffusion model M, Input audio A, Emotion Style S, Frame length N.
$L_A = length(A)$
$L_{exp} = length(S[0])$
Output $o = zeros(L_A, L_{exp})$
Condition $c = zeros(N, L_{exp} + 1)$
$c[:, 0, : -1] = S[random(1), :]$
$c[:, N/2 - 1 : N/2 + 2, : -1] = S[random(3), :]$
$c[:, 0, -1] = 1$
$c[:, N/2 - 1 : N/2 + 2, -1] = 1$
$i \leftarrow 0$
while $i < L_A$ **do**
    temp $= M$.sampling(cat($c, A[i : i + N]$))
    $c[:, 0, : -1] =$ temp$[-1]$
    $c[:, N/2 - 1 : N/2 + 2, : -1] = S[$random$(3), :]$
    $o[i : i + N] =$ temp
    $i \leftarrow i + N - 1$
end while
return o

tor with the last element set to 1. For style information, we select the middle three positions of the matrix and set them to the style values, with the last element also set to 1. We then combine this matrix with the audio feature frame by frame during training, which completes the setting of conditions. (See supplementary materials for details)

Long-term and dynamic sampling. During testing, we first select an emotional clip of a character from the dataset, such as "M003 Angry clip001", along with input audio. As shown in Alg. 1, to ensure continuity between sequences, we randomly select one frame as the initial state and subsequently use the last frame of the previous sequence as the initial state for the next sequence. This ensures the continuity of long sequences. To introduce diversity in each sequence generation, we randomly select 3 frames as the style each time, which allows for the generation of dynamic facial expressions within the overall sequence.

3.3. Lip Explicit Correlation Learning

After obtaining dynamic emotional expressions denoted as $\boldsymbol{x}_0$ from the diffusion model, we observed an unintended consequence in which the diffusion network inadvertently reduced the influence of audio, resulting in a noticeable misalignment between the audio and mouth shape. This phenomenon is likely attributed to the diffusion network's inherent inclination toward generating realistic sequences, which, in turn, diminishes the impact of the audio. To rectify this issue, we introduce lip-sync explicit correlation learning with refined mouth parameters optimization to capture acoustically aligned lip motion with expressive emotional talking faces. Our Lip-sync network incorporates an LSTM structure as the audio encoder and a CNN structure as the emotion encoder. Additionally, we utilized a lip-sync expert as a discriminator during the training of lip refinement. This design effectively generates mouth-related parameters that closely align with the input audio and emotional reference style. For a comprehensive understanding of our lip refinement network, we direct readers to the supplementary materials.

We use DeepSpeech [15] audio feature as a low-dimensional audio embedding. As shown in Fig. 3, our lip refinement framework contains two modules. First, we train an LSTM-based network Audio2BS which predicts neutral lip motion blendshape weights from input audio. This network is to learn the accurate talking motion closely associated with audio. To add emotional talking styles on the talking faces, we then introduce a new module Emotional-BS Residue Generator which predicts a blendshape weight residue between emotional talking lip motion and neutral lip motion. By adding the neutral talking face and the emotional BS, we can create an emotional talking face that is strongly correlated with the audio. Lastly, by integrating the talking face with the head pose, generated by the EmoDiff module, we use the lip part created by the lip refinement to replace the EmoDiff's result. This process ultimately yields an emotional talking face with enhanced lip-sync capabilities.

By following [32], we train a lip-sync expert. This expert directly captures lip features and audio features, facilitating similarity learning, which further constrains the correlation between audio and lip motion. $\mathcal{L}_{\text{sync}}$ is cosine-similarity with binary cross-entropy loss between audio features and lip features.

Subsequently, we employ these refined facial parameters and the generated head poses to animate a 3D blend shape rig. Utilizing GPU rendering, we obtain corresponding 3D rendered avatar images denoted as $I_{t,t\in\{0,..,N\}}$. Following this, we employ a video-to-video approach to generate talking face videos for arbitrary characters.

3.4. Face Neural Rendering

Upon acquiring images $I_{t,t\in\{0,..,N\}}$ from an external GPU renderer, we employ motion transfer techniques to achieve a realistic talking head effect for different subjects. Specifically, we utilize the Face-Vid2Vid method proposed by Wang et al [52] as the fundamental neural rendering pipeline $\mathcal{R}$. Furthermore, we conduct a fine-tuning process on the model using carefully selected high-resolution expressive talking videos from TalkHead-1HK dataset [52], aiming to enhance both expressiveness and rendering quality. In addition to fine-tuning, we augment the final image resolution to 512x512 using the face super-resolution method outlined in [2]. To ensure effective identity preservation throughout the process, we implement the relative mode technique developed by Siarohin et al. [35] for neural motion transfer. Specifically, we first render a refer-

Figure 4: Comparison of state-of-the-art models with our approach: In the first two comparisons, we conduct evaluations on the MEAD and HDTF datasets, respectively. For the third comparison, we utilize one AIGC-generated face. We also visualize our rig model results as intermediate representations. Our method consistently yields significantly superior results in terms of emotional expression, lip synchronization, identity preservation, and image quality. Please refer to our supplementary video for better comparison.

ence frame I_n with neural expression and then apply the relative motion $\mathcal{M}_{I_n \to I_t}$, which represents the transformation between talking frames and the neural frame, onto the source image T. Consequently, ultimate rendered outputs $\mathcal{R}(T, \mathcal{M}_{I_n \to I_t})$ is generated.

4. Experiments

4.1. Implementation Details

All experiments were conducted on a single V100 GPU utilizing the Adam optimizer [20]. The frame rate for all training datasets was set at 25 FPS. In the EmoDiff Module, our training primarily leveraged two datasets: the MEAD emotion dataset [51] and the HDTF multi-character dataset [59]. Each sequence generated during training consisted of a fixed length of 32 frames. We trained for a total of 1000 epochs with a batch size of 64 and a learning rate of 0.0004. For the Lip refinement model, we employed a sliding window of size $T = 8$ to extract training samples of audio features. The training process encompassed 50 epochs with a batch size of 32 and a learning rate of 0.0001.

Table 1: Comparison with state-of-the-art one-shot methods on MEAD and HDTF datasets. MakeitTalk and SadTalker maintain lip-sync and image quality without considering emotion. However, by adding emotions, EAMM and PD-FGC struggles with low image quality. Our method achieves both emotional expression and maintains lip-sync and image quality.

Method	MEAD					HDTF				
	LPIPS↓	CPBD↑	LMD↓	LSE-D↓	LSE-C↑	LPIPS↓	CPBD↑	LMD↓	LSE-D↓	LSE-C↑
Ground Truth	0	0.316	0	7.420	7.486	0	0.303	0	7.413	7.487
MakeitTalk [63]	0.295	0.213	4.178	10.151	5.012	0.289	0.247	5.026	10.334	4.823
SadTalker [58]	0.189	0.256	3.960	**9.634**	**6.095**	0.195	0.269	4.006	9.958	5.050
EAMM [18]	0.295	0.172	6.053	10.890	4.328	0.304	0.161	6.941	10.686	4.448
PD-FGC [48]	0.315	0.153	6.325	9.903	5.832	0.331	0.148	6.532	9.754	5.042
Ours	**0.169**	**0.299**	**3.845**	9.868	5.915	**0.176**	**0.280**	**3.948**	**9.233**	**5.263**

Figure 5: Comparison with the video sequences provided by EVP on emotion "angry". Since EVP requires separate training for each video, we cannot test it with arbitrary characters.

Table 2: Comparisons of state-of-the-art methods and our proposed method. SadTalker and MakeItTalk do not generate emotional speech. EAMM and PD-FGC produces emotional videos but loses identity and dynamic facial expressions. EVP is a video-based method that generates emotional speech but lacks emotion dynamics. In contrast, MagicTalk offers dynamic emotional expression with generated eye blinks and identity preservation.

Methods	Emotional talking	Generated eye blinks	Identity preservation
MakeItTalk [63]	✗	✗	✓
SadTalker [58]	✗	✓	✓
EAMM [18]	✓	✗	✗
PD-FGC [48]	✓	✗	✗
EVP [19]	✓	✗	✓
Ours	✓	✓	✓

4.2. Comparison with State-of-the-Arts

In Tab. 2, we offer an intuitive comparison of various methods' capabilities. It's clear that SadTalker and MakeItTalk lack the ability to generate emotional speech. Although EAMM can produce emotional videos, it does so at the expense of maintaining identity and fails to incorporate dynamic facial expressions, such as eye blinking. While PG-FGC can generate emotional speech, it has low image quality. EVP(Fig. 5) sacrifices the dynamism of emotions and cannot drive from a single image, limiting its applicability. In contrast, MagicTalk not only guarantees dynamic emotional expressions, such as generating high-frequency expressions like eye blinks, but also delivers high-quality videos with accurate lip-sync.

4.2.1 Qualitative Evaluation

As shown in Fig. 4, we compare our work with three state-of-the-art methods. MakeItTalk and SadTalker can't generate desired emotions from a single image. MakeItTalk's 2D approach lowers image quality, while SadTalker provides only neutral conversation. EAMM and PD-FGC generate emotional speech but sacrifices emotion dynamics and video quality. Our method excels in emotional expression, lip-sync, identity preservation, and image quality.

4.2.2 Quantitative Evaluation

Following established standards in the field, we utilize metrics for lip sync and image quality in our comparative analyses. For assessing the synchronization between lip movements and input audio, we employ SyncNet [7], which measures he distance score (LSE-D) and confidence score (LSE-C) to evaluate lip-sync precision. We also employ the Landmark Distance on the entire face (LMD) for a comprehensive evaluation of facial expression accuracy. For image quality, we assess image quality using learned perceptual image patch similarity (LPIPS) and cumulative probability blur detection (CPBD).

Our model exhibits enhanced performance over current leading methods, as demonstrated in Tab. 1. Comparing state-of-the-art methods on MEAD and HDTF datasets, MakeItTalk and SadTalker maintain lip-sync and image

Table 3: Computational complexity and parameter comparison of different methods. We report the GFLOPs and parameter counts for EAMM, PD-FGC, SadTalker, and our method.

	EAMM	PD-FGC	SadTalker	Ours
GFLOPs	57	230	762	641
Params	101M	146M	198M	168M

quality without considering emotion. However, adding emotion complicates lip-sync and rendering. EAMM faces challenges in achieving both emotion and maintaining lip-sync. PD-FGC has low image quality. Our method successfully balances emotional expression with lip-sync and image quality.

4.2.3 Computational Complexity and Parameters

We have calculated and compared our method's parameter count and computational complexity with the baseline models in Tab. 3. EAMM has a relatively small computational and parameter count but does not achieve high generation quality. The complexity of our method is similar to that of SadTalker.

4.3. Ablation Study

Our ablation study primarily investigates three key questions, including whether to employ an auto-regressive generation approach based on the initial state, whether to utilize a style encoding method with positional embedding, and whether to incorporate lip refinement for lip-sync results.

Emotional style embedding Unlike the approach used in EAMM, our emotional style employs positional embedding with a diffusion model. Here, we compare the use of a uniform emotional code against our method of positional embedding. It is observed that our diffusion method combined with positional embedding effectively captures high-frequency information. As illustrated in Fig. 6, the 3D rig is utilized to demonstrate the effects of employing Emotional style embedding. Our findings indicate that using our style embedding method effectively captures dynamic facial expression information. In contrast, when it is not utilized, maintaining effective emotional information proves to be challenging. We have also incorporated the Emotion-Fan metric [29] to evaluate emotional accuracy. The result highlights the superior performance of our approach, achieving an accuracy of 65.64 compared to 42.51 without Emotional Style Embedding. This further demonstrates the effectiveness of our method in preserving emotional expressiveness.

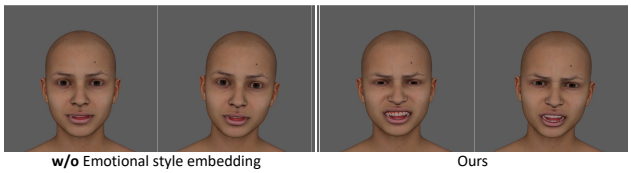

Figure 6: Ablation for emotional style embedding. Visualization of the 3D rig results in the 'angry' emotion. Without the emotion style embedding, it's challenging to maintain characteristics of anger, such as furrowing brows.

Lip explicit correlation learning While the diffusion network aids in generating dynamic emotions, we observed that it struggles to produce sequences that fully align with the audio. Hence, we employed a Lip Refinement Model to further optimize lip motion based on the audio. In Tab. 4, we measured the results using SyncNet and found Lip Refinement leads to more synchronized lip motion.

ARKit blend shapes Our ability to perform Lip Refinement is enabled by the choice of ARKit blend shapes. ARKit blend shapes allow us to separate mouth-related parameters, making it possible to refine lip movements independently without affecting other facial expressions. To validate this design choice, we also experimented with SyncNet-based optimization without separating mouth parameters. The results showed a Lip-sync confidence score of only 4.523, significantly lower than our method's 5.507. This highlights the effectiveness of using ARKit blend shapes for precise lip synchronization while maintaining overall facial expressiveness.

Autoregressive generation Since our goal is to generate indefinitely long sequences, we require an autoregressive method to achieve continuous talking results. In this approach, we use the last state of the previous sequence as a condition to generate the next sequence, thereby ensuring continuity between sequences. Refer to our supplementary for comparison.

4.4. User Study

Due to the subjective nature of emotion, quantitative evaluation is challenging. Therefore, we employed a subjective assessment method, involving 20 users who compared the results of different speech generation techniques. We provided reference images and emotional information for evaluation. The evaluation results depicted in Fig. 7 demonstrate that while EAMM and PD-FGC can generate emotion, it comes at the expense of video quality, leading to lower user ratings. Makeitalk and SadTalker, while lacking in emotion generation, achieved better overall quality than EAMM and PD-FGC. Our method, on the other hand, successfully maintains emotional intensity while attaining

Table 4: Ablation for lip explicit correlation learning. With lip refinement, the synchronization between the generated mouth and the audio improves, and mouth movements becomes closer to the ground truth (GT).

Method	LMD ↓	LSE-D ↓	LSE-C↑
w/o Lip refinement	4.167	10.583	4.338
Ours	**3.927**	**9.648**	**5.507**

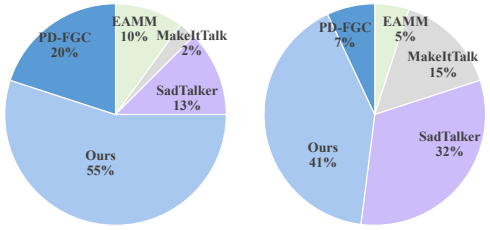

Figure 7: User study results show ratings for emotion preservation (on the left) and overall quality (on the right).

high-quality generation.

4.5. Applications

MagicTalk demonstrates its ability to generalize to various application scenarios, including facilitating dialogue generation for virtual characters using LLM, generating talking videos with different emotion styles, and animating real human faces or faces generated by AI-generated content (AIGC).

Virtual character dialogue generation with LLM Combining the powerful text generation capability of Large Language Models (LLM) with MagicTalk's robust animation ability, virtual dialogues are generated using LLM and converted into audio through Text-to-Speech. MagicTalk then utilizes the generated audio to produce dialogue videos. Furthermore, MagicTalk can incorporate emotion labels provided by LLM to create dialogues with specific emotions. This application has two notable characteristics: it can generate endless creative content and produce highly imaginative videos, such as the depiction of a quarrel between Mona Lisa and Leonardo da Vinci, as illustrated in Fig. 8.

Different emotions animation To further validate the universality and effectiveness of our method, we conducted experiments with various emotions. MagicTalk can animate different emotions for different individuals given the same audio input. As shown in Fig. 9, we animated three faces with various emotions, including Angry, Sad, Surprised, and Contempt.

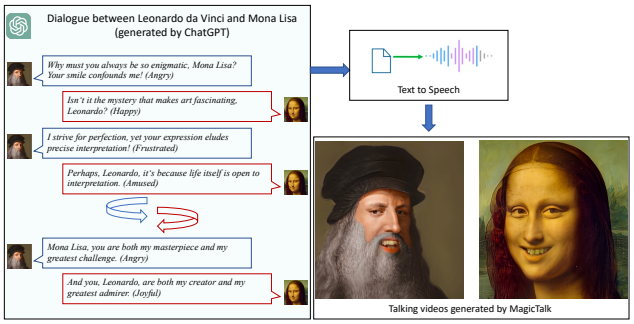

Figure 8: Virtual character dialogue generation. The virtual dialogue between Mona Lisa and Leonardo da Vinci was generated using ChatGPT and TTS. MagicTalk created corresponding emotional dialogue videos with input speech.

Figure 9: Different emotions animation. MagicTalk animates three faces with various emotions (Angry, Sad, Surprised, and Contempt) using the same audio input.

Real and AI-Generated faces animation MagicTalk exhibits strong generalization capabilities. Although trained on real human data, it can animate various types of images, including real human faces, portraits, and images generated by text-to-image models such as DALL·E3 [1]. As illustrated in Fig. 10, we first use DALL·E3 to generate face images based on text inputs. Then, MagicTalk can generate different talking videos based on arbitrary audio input.

5. Limitation and Social Impact

5.1. Limitations

Emotions in the dataset. We utilized the MEAD [51] dataset, which consists of artificially acted emotional expressions. Consequently, the emotional styles we generate tend to reflect acted emotions rather than the spontaneous flow of genuine emotions. Therefore, surpassing the limitations of these performed emotions in our generated emotional styles remains a challenge. A potential direc-

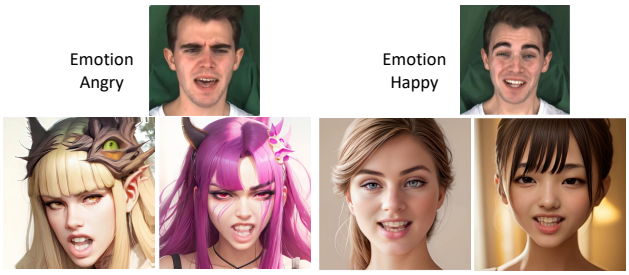

Figure 10: Real and AI-Generated faces animation. MagicTalk animates various image types, from real human faces to portraits and images created by text-to-image models like DALL·E3. Video results can be found on Project Page.

tion for future exploration is extracting emotions from real-life videos and incorporating them into our training process. This approach could lead to the creation of more natural and authentic emotional talking videos. We aim to enhance the emotional realism in our outputs, capturing the subtleties of true human emotions, thereby bridging the gap between artificial and genuine emotional expressions in synthesized video content.

5.2. Social Impact

With the widespread application of large-scale models, creating realistic talking videos has become extremely popular. There is significant demand for generating emotionally authentic talking videos. Our work is poised to have a profound impact on both the academic and industrial sectors. However, due to the potential for misuse of our technology in creating convincing talking videos, we restrict its use to academic purposes only. Additionally, any outputs generated using our technology should be clearly labeled as synthetic videos. On the other hand, our generation of emotional videos can also expand the dataset of synthetic videos, which will be beneficial for the further development of deep fake video detection.

6. Conclusion

In this paper, we present MagicTalk, a novel framework for generating emotionally expressive talking faces with precise lip synchronization. Our approach employs collaborative implicit-explicit correlation learning to model implicit and explicit relationships between audio inputs and emotional talking heads. By integrating audio conditions with emotional style embeddings into the diffusion process, we capture the implicit correlation between audio and emotional expressions. Furthermore, our lip-sync explicit correlation learning, along with refined mouth parameter optimization, ensures acoustically aligned lip motion with expressive emotional talking faces. Our method outperforms existing techniques, demonstrating improved facial emotional expressiveness while maintaining high video quality. MagicTalk represents a significant advancement in emotional talking face generation, enabling the creation of realistic and emotionally engaging digital human representations across various applications.

References

[1] J. Betker, G. Goh, L. Jing, T. Brooks, J. Wang, L. Li, L. Ouyang, J. Zhuang, J. Lee, Y. Guo, et al. Improving image generation with better captions. 2023. 10

[2] K. C. Chan, X. Wang, X. Xu, J. Gu, and C. C. Loy. Glean: Generative latent bank for large-factor image super-resolution. In *CVPR*, 2021. 6

[3] L. Chen, G. Cui, C. Liu, Z. Li, Z. Kou, Y. Xu, and C. Xu. Talking-head generation with rhythmic head motion. In *ECCV*, 2020. 3

[4] L. Chen, R. K. Maddox, Z. Duan, and C. Xu. Hierarchical cross-modal talking face generation with dynamic pixel-wise loss. In *CVPR*, 2019. 3

[5] J. S. Chung, A. Jamaludin, and A. Zisserman. You said that? *arXiv*, 2017. 3

[6] J. S. Chung, A. Nagrani, and A. Zisserman. VoxCeleb2: Deep Speaker Recognition. In *Proceedings of Interspeech*, 2018. 2

[7] J. S. Chung and A. Zisserman. Out of time: automated lip sync in the wild. In *ACCV*, 2016. 8

[8] D. Das, S. Biswas, S. Sinha, and B. Bhowmick. Speech-driven facial animation using cascaded gans for learning of motion and texture. In *ECCV*, 2020. 3

[9] H. Ding, K. Sricharan, and R. Chellappa. ExprGAN: Facial expression editing with controllable expression intensity. In *AAAI*, 2018. 3

[10] S. E. Eskimez, Y. Zhang, and Z. Duan. Speech driven talking face generation from a single image and an emotion condition. *IEEE TMM*, 2021. 3

[11] Y. Gan, Z. Yang, X. Yue, L. Sun, and Y. Yang. Efficient emotional adaptation for audio-driven talking-head generation. In *ICCV*, 2023. 3

[12] J. Guan, Z. Zhang, H. Zhou, T. Hu, K. Wang, D. He, H. Feng, J. Liu, E. Ding, Z. Liu, et al. Stylesync: High-fidelity generalized and personalized lip sync in style-based generator. In *CVPR*, 2023. 3

[13] Y. Guo, K. Chen, S. Liang, Y.-J. Liu, H. Bao, and J. Zhang. AD-NERF: Audio driven neural radiance fields for talking head synthesis. In *ICCV*, 2021. 3

[14] S. Gururani, A. Mallya, T.-C. Wang, R. Valle, and M.-Y. Liu. SPACE: Speech-driven portrait animation with controllable expression. *arXiv*, 2022. 2, 4

[15] A. Hannun, C. Case, J. Casper, B. Catanzaro, G. Diamos, E. Elsen, R. Prenger, S. Satheesh, S. Sengupta, A. Coates, et al. Deep speech: Scaling up end-to-end speech recognition. *arXiv*, 2014. 6

[16] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. *NeurIPS*, 2020. 4

[17] J. Ho and T. Salimans. Classifier-free diffusion guidance. *arXiv*, 2022. 5

[18] X. Ji, H. Zhou, K. Wang, Q. Wu, W. Wu, F. Xu, and X. Cao. EAMM: One-shot emotional talking face via audio-based emotion-aware motion model. In *SIGGRAPH Conference*, 2022. 2, 3, 4, 8

[19] X. Ji, H. Zhou, K. Wang, W. Wu, C. C. Loy, X. Cao, and F. Xu. Audio-driven emotional video portraits. In *CVPR*, 2021. 2, 4, 8

[20] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 7

[21] A. Lahiri, V. Kwatra, C. Frueh, J. Lewis, and C. Bregler. Lipsync3d: Data-efficient learning of personalized 3d talking faces from video using pose and lighting normalization. In *CVPR*, 2021. 3

[22] L. Li, S. Wang, Z. Zhang, Y. Ding, Y. Zheng, X. Yu, and C. Fan. Write-a-speaker: Text-based emotional and rhythmic talking-head generation. In *AAAI*, 2021. 3, 4

[23] T. Li, T. Bolkart, M. J. Black, H. Li, and J. Romero. Learning a model of facial shape and expression from 4d scans. *ACM TOG*, 2017. 2

[24] B. Liang, Y. Pan, Z. Guo, H. Zhou, Z. Hong, X. Han, J. Han, J. Liu, E. Ding, and J. Wang. Expressive talking head generation with granular audio-visual control. In *CVPR*, 2022. 3, 4

[25] H. Liu, Z. Zhu, N. Iwamoto, Y. Peng, Z. Li, Y. Zhou, E. Bozkurt, and B. Zheng. Beat: A large-scale semantic and emotional multi-modal dataset for conversational gestures synthesis. In *ECCV*, 2022. 2, 4

[26] X. Liu, Y. Xu, Q. Wu, H. Zhou, W. Wu, and B. Zhou. Semantic-aware implicit neural audio-driven video portrait generation. In *ECCV*, 2022. 3

[27] S. R. Livingstone and F. A. Russo. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in north american english. 2018. 2

[28] Y. Ma, S. Wang, Z. Hu, C. Fan, T. Lv, Y. Ding, Z. Deng, and X. Yu. Styletalk: One-shot talking head generation with controllable speaking styles. *arXiv*, 2023. 2

[29] D. Meng, X. Peng, K. Wang, and Y. Qiao. Frame attention networks for facial expression recognition in videos. In *ICIP*, 2019. 9

[30] M. Meshry, S. Suri, L. S. Davis, and A. Shrivastava. Learned spatial representations for few-shot talking-head synthesis. In *ICCV*, 2021. 3

[31] Z. Peng, H. Wu, Z. Song, H. Xu, X. Zhu, J. He, H. Liu, and Z. Fan. Emotalk: Speech-driven emotional disentanglement for 3d face animation. In *ICCV*, 2023. 3

[32] K. Prajwal, R. Mukhopadhyay, V. P. Namboodiri, and C. Jawahar. A lip sync expert is all you need for speech to lip generation in the wild. In *ACM MM*, 2020. 6

[33] S. Shen, W. Li, Z. Zhu, Y. Duan, J. Zhou, and J. Lu. Learning dynamic facial radiance fields for few-shot talking head synthesis. In *ECCV*, 2022. 3

[34] S. Shen, W. Zhao, Z. Meng, W. Li, Z. Zhu, J. Zhou, and J. Lu. Difftalk: Crafting diffusion models for generalized audio-driven portraits animation. In *CVPR*, 2023. 3

[35] A. Siarohin, S. Lathuilière, S. Tulyakov, E. Ricci, and N. Sebe. First order motion model for image animation. In *NeurIPS*, 2019. 6

[36] S. Sinha, S. Biswas, R. Yadav, and B. Bhowmick. Emotion-controllable generalized talking face generation. *arXiv*, 2022. 4

[37] L. Song, W. Wu, C. Qian, R. He, and C. C. Loy. Everybody's talkin': Let me talk as you want. *IEEE Transactions on Information Forensics and Security*, 2022. 3

[38] M. Stypułkowski, K. Vougioukas, S. He, M. Zięba, S. Petridis, and M. Pantic. Diffused heads: Diffusion models beat gans on talking-face generation. *arXiv*, 2023. 3

[39] Y. Sun, H. Zhou, K. Wang, Q. Wu, Z. Hong, J. Liu, E. Ding, J. Wang, Z. Liu, and K. Hideki. Masked lip-sync prediction by audio-visual contextual exploitation in transformers. In *SIGGRAPH Asia*, 2022. 3

[40] S. Suwajanakorn, S. M. Seitz, and I. Kemelmacher-Shlizerman. Synthesizing obama: Learning lip sync from audio. *ACM TOG*, 2017. 3

[41] S. Tan, B. Ji, and Y. Pan. Emmn: Emotional motion memory network for audio-driven emotional talking face generation. In *ICCV*, 2023. 3, 4

[42] A. Tang, T. He, X. Tan, J. Ling, R. Li, S. Zhao, L. Song, and J. Bian. Memories are one-to-many mapping alleviators in talking face generation. *arXiv preprint*, 2022. 3

[43] J. Tang, K. Wang, H. Zhou, X. Chen, D. He, T. Hu, J. Liu, G. Zeng, and J. Wang. Real-time neural radiance talking portrait synthesis via audio-spatial decomposition. *arXiv preprint*, 2022. 3

[44] S. Taylor, T. Kim, Y. Yue, M. Mahler, J. Krahe, A. G. Rodriguez, J. Hodgins, and I. Matthews. A deep learning approach for generalized speech animation. *ACM TOG*, 2017. 3

[45] J. Thies, M. Elgharib, A. Tewari, C. Theobalt, and M. Nießner. Neural voice puppetry: Audio-driven facial reenactment. In *ECCV*, 2020. 3

[46] L. Tian, Q. Wang, B. Zhang, and L. Bo. Emo: Emote portrait alive-generating expressive portrait videos with audio2video diffusion model under weak conditions. *arXiv preprint*, 2024. 3

[47] J. Tseng, R. Castellon, and C. K. Liu. EDGE: Editable dance generation from music. *arXiv*, 2022. 5

[48] D. Wang, Y. Deng, Z. Yin, H.-Y. Shum, and B. Wang. Progressive disentangled representation learning for fine-grained controllable talking head synthesis. In *CVPR*, 2023. 3, 8

[49] J. Wang, K. Zhao, Y. Ma, S. Zhang, Y. Zhang, Y. Shen, D. Zhao, and J. Zhou. Facecomposer: A unified model for versatile facial content creation. *NeurIPS*, 2024. 3

[50] J. Wang, K. Zhao, S. Zhang, Y. Zhang, Y. Shen, D. Zhao, and J. Zhou. Lipformer: High-fidelity and generalizable talking face generation with a pre-learned facial codebook. In *CVPR*, 2023. 3

[51] K. Wang, Q. Wu, L. Song, Z. Yang, W. Wu, C. Qian, R. He, Y. Qiao, and C. C. Loy. MEAD: A large-scale audio-visual dataset for emotional talking-face generation. In *ECCV*, 2020. 2, 4, 7, 10

[52] T.-C. Wang, A. Mallya, and M.-Y. Liu. One-shot free-view neural talking-head synthesis for video conferencing. In *CVPR*, 2021. 3, 6

[53] C. Xu, J. Zhu, J. Zhang, Y. Han, W. Chu, Y. Tai, C. Wang, Z. Xie, and Y. Liu. High-fidelity generalized emotional talking face generation with multi-modal emotion space learning. In *CVPR*, 2023. 3

[54] Z. Ye, Z. Jiang, Y. Ren, J. Liu, J. He, and Z. Zhao. Geneface: Generalized and high-fidelity audio-driven 3D talking face synthesis. *arXiv*, 2023. 3

[55] R. Yi, Z. Ye, J. Zhang, H. Bao, and Y.-J. Liu. Audio-driven talking face video generation with learning-based personalized head pose. *arXiv preprint*, 2020. 3

[56] C. Zhang, S. Ni, Z. Fan, H. Li, M. Zeng, M. Budagavi, and X. Guo. 3d talking face with personalized pose dynamics. *IEEE TVCG*, 2021. 3

[57] C. Zhang, Y. Zhao, Y. Huang, M. Zeng, S. Ni, M. Budagavi, and X. Guo. FACIAL: Synthesizing dynamic talking face with implicit attribute learning. In *ICCV*, 2021. 3

[58] W. Zhang, X. Cun, X. Wang, Y. Zhang, X. Shen, Y. Guo, Y. Shan, and F. Wang. Sadtalker: Learning realistic 3d motion coefficients for stylized audio-driven single image talking face animation. In *CVPR*, 2023. 3, 8

[59] Z. Zhang, L. Li, Y. Ding, and C. Fan. Flow-guided one-shot talking face generation with a high-resolution audio-visual dataset. In *CVPR*, 2021. 3, 7

[60] W. Zhong, C. Fang, Y. Cai, P. Wei, G. Zhao, L. Lin, and G. Li. Identity-preserving talking face generation with landmark and appearance priors. *arXiv*, 2023. 3

[61] H. Zhou, Y. Liu, Z. Liu, P. Luo, and X. Wang. Talking face generation by adversarially disentangled audio-visual representation. In *AAAI*, 2019. 3

[62] H. Zhou, Y. Sun, W. Wu, C. C. Loy, X. Wang, and Z. Liu. Pose-controllable talking face generation by implicitly modularized audio-visual representation. In *CVPR*, 2021. 3

[63] Y. Zhou, X. Han, E. Shechtman, J. Echevarria, E. Kalogerakis, and D. Li. MakeItTalk: Speaker-aware talking-head animation. *ACM TOG*, 2020. 3, 4, 8

[64] H. Zhu, H. Huang, Y. Li, A. Zheng, and R. He. Arbitrary talking face generation via attentional audio-visual coherence learning. *arXiv preprint*, 2018. 3

[65] L. Zhu, X. Liu, X. Liu, R. Qian, Z. Liu, and L. Yu. Taming diffusion models for audio-driven co-speech gesture generation. *arXiv*, 2023. 4, 5