Parameter Efficient Emotional Transfer for Fine-Grained Emotional Audio-Driven Talking Portrait Generation

Yiguo Jiang, Xiaodong Cun, *Member, IEEE*, Yong Zhang, Tao He, Fan Tang, and Chi-Man Pun

***Abstract*—Facial expression, especially emotion, plays a crucial role in building a realistic talking portrait video from audio. Previous emotional talking head methods train lip-synchronization with video-level annotated emotion labels jointly. However, the video-level annotated emotion labels make it hard to learn the accurate frame-wise emotions. Besides, since the emotional dataset is much smaller than the neutral faces, it becomes harder to generate lip-synchronized portrait videos with accurate expressions. To address these challenges, we propose a new method for efficient emotional transfer and detailed emotion capture. In detail, we first consider emotions as the residual of the neutral expression and train additional adapters (LoRAs in our case) for each emotion individually. Then, we use the frame-wise emotional annotations as input and design a novel method to jointly optimize the multiple adapters. After training, our method can produce high visual quality results by driving audio and fine-grained emotion guidance with negligible additional parameters. Extensive experiments also demonstrate the effectiveness of theproposed approach in synthesizing photo-realistic, emotionally expressive talking portraits that are well-synchronized with the input audio compared with other state-of-the-art methods.**

***Index Terms*—Fine-grained emotional expression, parameter efficiency, talking portrait generation.**

I. Introduction

AUDIO-DRIVEN talking portrait generation aims to create realistic 2D portrait videos from audio, which is a crucial task in computer graphics and multimedia. Such technology has a wide range of applications, including the creation of virtual avatars, dubbing for media, and enabling telepresence communications. In addition to the synchronization with audio, the emotion of the generated portrait is also a necessary component that contributes to the vividness and realism of facial animations, as it significantly affects the perceived authenticity of the generated content.

Previous works [1], [2], [3], [4], [5], [6], [7], [8], have largely centered on audio-lip synchronization, often overlooking the aspect of controlling facial emotions in the generated videos. Some recent methods [9], [10], [11], [12], [13], [14], [15], [16], [17], [18] have integrated emotion control into talking portrait generation to create expressive emotional talking heads. Although these emotional audio-driven talking head generation techniques have advanced rapidly, two primary issues persist. First, these methods typically incur high costs [10], [13], [14], [15], [16] due to the need for training or fine-tuning the entire network for emotional talking heads generation. Second, the quest for more precise, vivid, and fine-grained emotional control is an ongoing challenge. We argue that these limitations might be attributed to two main factors. On the one hand, the emotional dataset is limited compared with the talking head videos of the neutral face, and there is generally an insufficient emphasis on the efficient reuse of knowledge acquired from emotion-agnostic datasets. On the other hand, the utilization of emotion labels at the video-level [9], [11], [12], [15] instead of frame-by-frame leads to a generalized, mean emotive representation of the video, rather than capturing the exact emotions conveyed in each frame.

To solve these challenges, we propose an efficient and effective emotional transfer method by leveraging the fine-grained expression information encoded in each video frame. The parameter-efficient transfer learning [19] serves as a good way to relieve the requirements of the emotional dataset. Furthermore, employing frame-wise emotional guidance as conditional information is also important to achieve fine-grained control of each frame. Frame-wise guidance can provide a more detailed representation of expressive features in each frame compared to static video-level emotional labels.

Specifically, to implement the paradigm mentioned above, we propose an audio-to-mesh and then a mesh-to-video pipeline as shown in Fig. 2. In detail, we begin with an audio-to-3D mesh model trained on datasets with neutral expressions to focus on audio-lip synchronization, which subsequently allows

Received 2 July 2025; revised 27 October 2025 and 11 December 2025; accepted 13 January 2026. Date of publication 1 June 2026; date of current version 8 October 2026. This work was supported in part by Science and Technology Development Fund, Macau SAR, under Grant 0193/2023/RIA3 and Grant 0079/2025/AFJ, in part by the University of Macau under Grant MYRG-GRG2024-00065-FST-UMDF, in part by Guangdong Provincial Regional Joint Fund - Young Scientists Fund under Project 2024A1515110052, and in part by the Young Scientists Fund of the Natural Science Foundation of China under Project 62506064. The associate editor coordinating the review of this article and approving it for publication was Prof. Jungong Han. *(Corresponding authors: Chi-Man Pun; Xiaodong Cun.)*

Yiguo Jiang and Chi-Man Pun are with the Department of Computer and Information Science, University of Macau, Macau 999078, China (e-mail: yc27960@umac.mo; cmpun@umac.mo).

Xiaodong Cun is with the School of Computing and Information Technology, Great Bay University, Dongguan 523000, China (e-mail: vinthony@gmail.com).

Yong Zhang is with Meituan, Beijing 100102, China (e-mail: zhangyong201303@gmail.com).

Tao He is with GRG Banking Equipment Company, Ltd., Guangzhou 510663, China (e-mail: kevin.92.he@gmail.com).

Fan Tang is with the Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100864, China (e-mail: tangfan@ict.ac.cn).

This article has supplementary downloadable material available at https://doi.org/10.1109/TMM.2026.3699022, provided by the authors.

Digital Object Identifier 10.1109/TMM.2026.3699022

1520-9210 © 2026 IEEE. All rights reserved, including rights for text and data mining, and training of artificial intelligence and similar technologies.
Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

Authorized licensed use limited to: INSTITUTE OF COMPUTING TECHNOLOGY CAS. Downloaded on October 19,2026 at 08:40:51 UTC from IEEE Xplore. Restrictions apply.

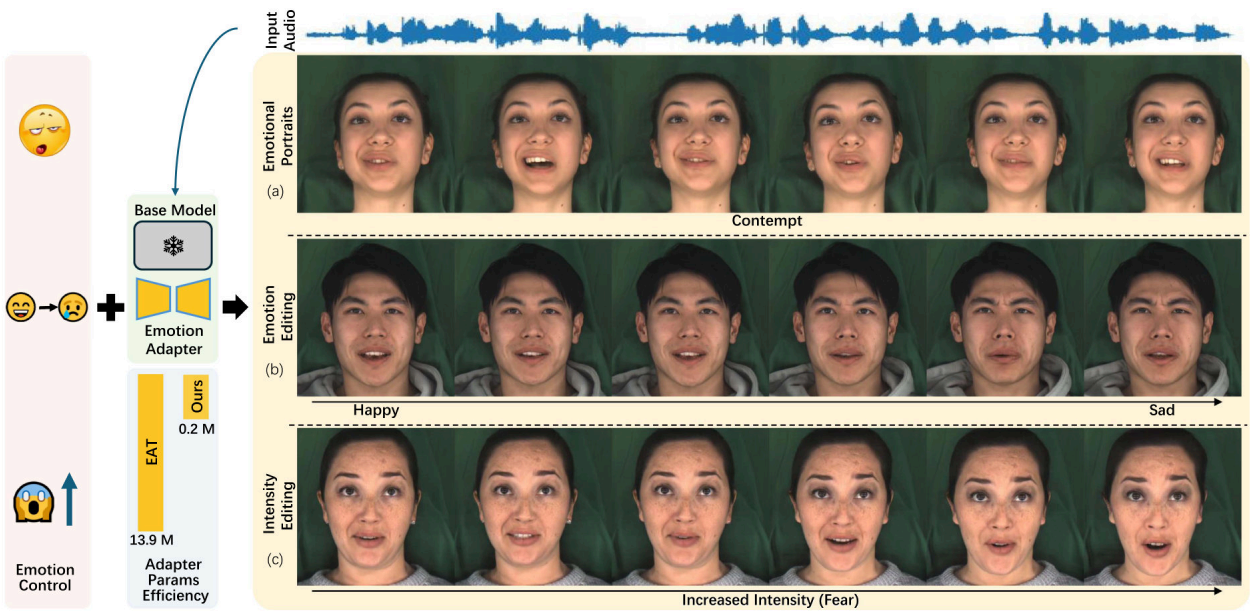

Fig. 1. Our method enables flexible emotional control for audio-driven talking-head generation by leveraging an efficient emotional adapter. Given a driving audio clip and the emotion label, our approach faithfully synthesizes expressive talking-head portraits with the target emotion via a **0.2M-parameter** emotional adapter. We adopt fine-grained, frame-wise emotional guidance to maintain the temporal continuity of learned emotional features, which further enables editing of the generated videos, including modification of the emotion category and its intensity. Here, (a) shows inputting an audio clip to drive people to speak with contempt emotion. (b) shows that emotions can transition smoothly between happy and sad **throughout** the video. (c) shows the intensity of the fear emotion increases continuously **throughout** the video.

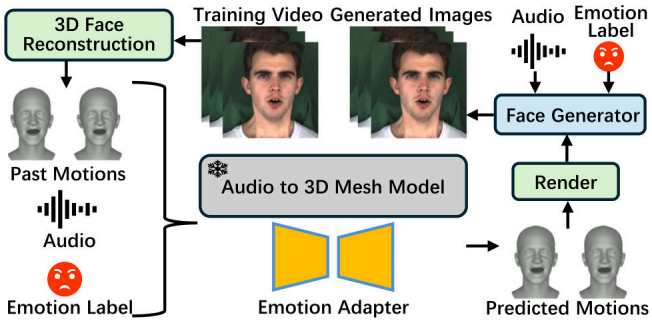

Fig. 2. The framework of our proposed method mainly consists of the following parts. We initially train an emotion-agnostic base model for mapping audio inputs to 3D mesh representations, as described in Section III-B1. Subsequently, in Sections III-B2 and III-B3, we develop an emotion adapter to facilitate emotion manipulation, building upon the aforementioned base model. Lastly, we employ a face generator to transform the emotional 3D meshes into photorealistic expressive 2D portrait images to synthesize output video in Section III-C.

for the use of a smaller emotional dataset in developing a novel method for transferring emotion-related knowledge using emotional talking videos. Building upon the audio-to-3D mesh base model trained on neutral expressions, we use a Low-Rank Adaptation (LoRA) [19] to facilitate the transfer of emotional knowledge. Besides, to address the issue that prior emotional talking head generation methods rely solely on video-level one-hot emotional labels for emotion control, we propose a novel adapter fusion method of the emotion-specific adapters, based on the expression information of each video frame. Our merge method enables the adapter to learn the expression information more accurately through the frame-wise emotional feature during the fusion process, thereby enabling the merged adapter to achieve enhanced precision in emotional control. Through the above steps, our method can achieve fine-grained emotion control with negligible additional parameters in audio-to-3D mesh mapping. Finally, an audio-guided face generator is used to synthesize the 3D meshes into photo-realistic 2D emotional portraits. Experiments show that the results of the proposed method achieve state-of-the-art performance compared to the baseline methods in terms of visual quality, lip-synchronization, and emotion accuracy.

In summary, our contributions are summarized as:

- We propose a novel method for fine-grained emotional talking portrait generation by an efficient emotion transfer mechanism based on LoRA adapters, which allows the model to be quickly adapted to new emotional domains without catastrophic forgetting.
- We present a novel adapter combination method that leverages frame-wise emotional label to fuse the individual emotion-specific adapters to learn more precise expression features.
- Extensive experiments demonstrate the effectiveness of our proposed approach in synthesizing photorealistic, emotionally expressive talking portraits that are well-synchronized with the input audio.

II. Related Work

A. Audio-Driven Talking Portrait Generation

Audio-driven talking portrait generation aims to produce realistic talking portrait videos that are synchronized with the corresponding speech content. Recent advances in deep learning-based approaches [1], [2], [3], [4], [5], [6], [7], [8], [20], [21], [22], [23], [24], [25], [26] have significantly progressed this field. Some methods directly model the mapping between speech and

Authorized licensed use limited to: INSTITUTE OF COMPUTING TECHNOLOGY CAS. Downloaded on October 19,2026 at 08:40:51 UTC from IEEE Xplore. Restrictions apply.

facial features [1], [6], [27], [28]. Others use intermediate representations, such as facial landmarks [20], [29], [30] and parameters of a 3D Morphable Model (3DMM) [2], [21], [31]. For example, Wav2Lip [1] demonstrates effective synchronization between lip motion and audio by employing an expert SyncNet [32] to align the speech and mouth region. MakeItTalk [20] predicts landmarks from disentangled audio content and speaker information. SadTalker [2] takes the approach of learning the expression and pose coefficients of a 3DMM separately. Aniportrait [22] first extracts a sequence of landmarks and then uses this landmark sequence to generate portrait videos with temporal stability. Hallo [6] incorporates a hierarchical audio-driven visual synthesis module to improve the precision of audio-visual alignment. EchoMimic [8] generates high-fidelity talking head videos by leveraging audio input and facial landmarks. Despite these advancements, the majority of these methods primarily focus on generating synchronized motion in the mouth region. None of these existing methods take into account emotional control.

B. Emotional Audio-Driven Talking Portrait

Emotion enhances the vividness and expressiveness of facial animation, so it significantly affects the realism of the generated portrait. Recently, some audio-driven talking portrait generation methods have been focused on expression control in generating portraits [9], [10], [11], [12], [13], [14], [15], [16], [17], [18], [33], [34]. MEAD [9] introduces a high-quality talking head video dataset annotated with eight emotion categories and proposes a method to condition talking head generation using a one-hot emotion label. EVP [10] decomposes input audio into disentangled content and emotion embeddings. EAMM [13] generates emotional talking faces by extracting dynamic emotional patterns from a source video and applying these transferable patterns to an arbitrary audio-driven talking face. PD-FGC [14] represents different facial motion factors via disentangled latent representations and leverages an image generator to synthesize talking heads from these latent representations. EAT [17] presents an approach to transform emotion-agnostic talking-head models into emotion-controllable ones to increase the flexibility of the method. EDTalk [18] employs three modules to decompose the facial dynamics into three distinct latent spaces to represent the mouth, pose, and expression, respectively. Meanwhile, there are some 3D emotional audio-driven talking head generation methods [35], [36]. EmoTalk [35] disentangles different emotions in the speech to generate rich 3D facial expressions. EMOTE [36] proposes a supervision mechanism with perceptual lip-reading and dynamic emotion losses and a content-emotion disentanglement mechanism. These methods exhibit two primary limitations. First, training the entire model to achieve emotional talking-head generation is inefficient. Second, using video-level emotional labels for supervision tends to overlook the fine-grained expressive features present in each frame. In our work, we achieve efficient emotional transfer by adapting the emotional LoRA adapters to a basic neutral talking-head model. We then use frame-level emotion guidance to fuse these adapters, thereby enabling the model to learn fine-grained expressive features.

C. LoRA

LoRA (Low-Rank Adaptation) [19] is an efficient technique for adapting and fine-tuning pre-trained large language models (LLMs) with a minimal increase in the number of trainable parameters. Given a frozen pre-trained weight matrix $W_0 \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$, LoRA introduces a low-rank update $\Delta W = BA$ to capture the fine-tuning adjustments to the original weights, where $B \in \mathbb{R}^{d_{\text{out}} \times r}$ and $A \in \mathbb{R}^{r \times d_{\text{in}}}$ with $r \ll \min(d_{\text{out}}, d_{\text{in}})$. Instead of updating the entire weight matrix W_0 directly, LoRA trains only these compact low-rank matrices A and B, thereby accelerating the adaptation process and diminishing computational expenses. Formally, LoRA can be represented as:

$$y = (W_0 + \Delta W)x = W_0 x + BAx. \tag{1}$$

D. LoRA Combinations

In our task and other multiple LoRA tasks, model merging techniques have emerged to combine multiple task-specific models into a single model with diverse abilities. Specifically, assume there are K task-specific LoRA adapters for the same base layer. Each adapter is represented by an update $\Delta W_k = B_k A_k$ with rank r_k. The goal is to construct a merged update ΔW^* that integrates these capabilities within a target rank and with minimal interference. Such merging techniques include concatenation (Cat) [37], Singular Value Decomposition (SVD) [37], Task Arithmetic [38], DARE [39], and TIES [40]. Cat combination [37] concatenates the K LoRA matrices along the rank dimension as follows: $A_{merged} = concat(A_1, \ldots, A_K, dim = 0)$, $B_{merged} = concat(B_1, \ldots, B_K, dim = 1)$. So this yields $\Delta W^* = B_{merged} A_{merged} = \sum_k B_k A_k$. This approach is simple but the resulting merged adapter can have a large size due to concatenation if many adapters are combined. Based on the Cat results, SVD combination [37] applies singular value decomposition to enforce a rank budget. It can capture the principal shared directions but compromise task-specific LoRA adapters. Task Arithmetic [38] forms a weighted sum of K task-specific LoRA matrices as follows: $A_{merged} = \sum_k weight_k * A_k$, $B_{merged} = \sum_k weight_k * B_k$. DARE [39] aggregates K task-specific LoRA matrices by randomly dropping parameters at a specified drop rate and rescaling the retained parameters. TIES [40] is a three-stage algorithm for model merging. It proceeds as follows: first, redundant parameters are pruned; second, sign conflicts are resolved by forming an aggregate vector; and third, parameters whose signs match the aggregate are averaged. This procedure mitigates performance degradation arising from redundancy and sign disagreement in the merged model. These adapter-merging methods often fail to fully preserve the knowledge encapsulated in individual adapters, rendering them unsuitable for direct application to our approach. In our work, we extend the TIES combination strategy to integrate emotion-specific adapters and use fine-grained expression features encoded in individual video frames to further refine the resulting merged adapter.

Authorized licensed use limited to: INSTITUTE OF COMPUTING TECHNOLOGY CAS. Downloaded on October 19,2026 at 08:40:51 UTC from IEEE Xplore. Restrictions apply.

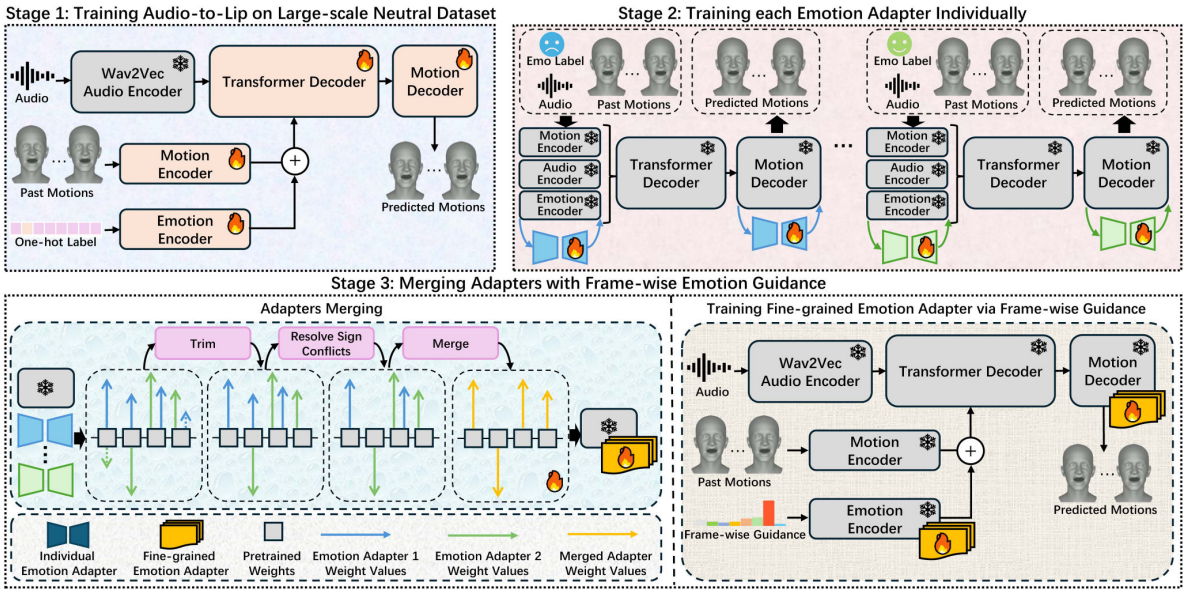

Fig. 3. The pipeline of our audio-to-3D mesh method. We achieve audio-driven emotional 3D mesh generation in three stages. In Stage 1 (Section III-B1), the basic model is trained for audio-to-3D mesh conversion using a dataset of neutral expressions. For Stage 2 (Section III-B2), we train the emotion-specific adapter for each emotion, leveraging the basic model from Stage 1. In Stage 3 (Section III-B3), these individual emotion-specific adapters are integrated to form a cohesive representation, which is then further refined with fine-grained emotional guidance from each frame.

III. Methodology

A. Overview

We aim to generate high-fidelity emotional talking portraits from audio. As shown in Fig. 2, our method first produces high-quality 3D emotional portraits from the audio with emotion labels. We leverage the knowledge from a basic audio-to-3D mesh model on the emotion-agnostic datasets, and then the new emotions are transferred via the emotional LoRA adapters. After that, we combine the LoRA adapters from different emotions and retrain them on the fine-grained emotion guidance to facilitate the generation of emotional 3D meshes. Finally, the high-quality emotional 2D video portraits are produced from the generated 3D-driven signals by our face render generator. In the following sections, we give the details.

B. Audio to 3D Mesh Generation

Given a driving audio clip $\mathbf{a}$ and the emotion label $\mathbf{e_n}$, the goal here is to create an audio-to-3D mesh model $\mathbf{M}$ to generate the emotional 3D mesh sequence $\hat{\mathbf{Y}}_\mathbf{t} = (\hat{\mathbf{y}}_\mathbf{1}, \ldots, \hat{\mathbf{y}}_\mathbf{T})$, where $\mathbf{T}$ is the length of visual frame sequence and $\mathbf{n}$ represents n kinds of emotions. So conditioned on driving audio $\mathbf{a}$ and emotion label $\mathbf{e_n}$, our audio-to-3D mesh model can predict emotional facial motion based on past facial motions, formulated as:

$$\hat{\mathbf{y}}_\mathbf{t} = \mathbf{M}_\theta(\mathbf{a}, \mathbf{e_n}, \hat{\mathbf{y}}_{<\mathbf{t}}), \tag{2}$$

where θ represents the parameters of our audio-to-3D mesh model.

Specifically, our emotional 3D mesh generation method contains three stages as shown in Fig. 3. In stage one, we train the basic audio-to-3D mesh model with the neutral expression. Then, we train each emotion-specific adapter to achieve efficient emotion transfer. Finally, these emotion adapters are merged using frame-wise emotional guidance. Below, we give the details of each stage.

1) Training the Audio to 3D Mesh Model on Neutral Videos: By considering the emotional talking head model as the offset of the emotion-agnostic model, we first train a base model on neutral expression to obtain an accurate generic audio-to-3D mesh model. The details of the model architecture are shown in stage 1 of Fig. 3. Following the state-of-the-art audio-to-3D mesh generation model, i.e., FaceFormer [41], we use the pre-trained speech model wav2vec 2.0 [42] as the audio encoder $\mathbf{E_{audio}}$, which converts the raw waveform speech into contextualized speech representations. A linear layer serves as our motion encoder $\mathbf{E_{motion}}$, projecting the input face motion into the model's dimensional space. Then, we employ a masked transformer decoder $\mathbf{D}$ similar to the Faceformer [41] to process the contextual speech representation and past facial motion features, thereby generating the predicted facial motions. This transformer decoder $\mathbf{D}$ uses causal self-attention to model the dependencies between each frame in the context of past motions, and cross-modal attention to align the audio and motion modalities. Subsequently, a linear motion decoder $\mathbf{D_{motion}}$ maps the predicted facial motions to 3D meshes. In this stage, the audio-to-3D mesh model is trained on the neutral face dataset [43], which have minimal facial emotion but only have movement of the mouth area. Consequently, the emotion label input to the emotion encoder $\mathbf{E_{emotion}}$ does not contain expression information. In summary, this part of the training is shown in stage 1 of Fig. 3, which can be expressed as:

$$\hat{\mathbf{y}}_\mathbf{t} = \mathbf{D_{motion}}(\mathbf{D}(\mathbf{E_{audio}}(\mathbf{a}), \mathbf{E_{emotion}}(\mathbf{e_n}) + \mathbf{E_{motion}}(\hat{\mathbf{y}}_{<\mathbf{t}}))). \tag{3}$$

Authorized licensed use limited to: INSTITUTE OF COMPUTING TECHNOLOGY CAS. Downloaded on October 19,2026 at 08:40:51 UTC from IEEE Xplore. Restrictions apply.

2) Training Individual Emotion Adapter Using LoRA: After training on the neutral expression, we aim to learn the specific emotional talking head based on the neutral model using the emotional talking head dataset. As shown in stage 2 of Fig. 3, we freeze the parameters in the base model and train emotion-specific LoRA adapters to efficiently and effectively transfer emotional knowledge without excessive computational costs. Our audio-to-3D base model comprises various components, such as the MLP emotion encoder $\mathbf{E_{emotion}}$, MLP motion decoder $\mathbf{D_{motion}}$, and the attention module within the transformer decoder $\mathbf{D}$. However, we find that adding LoRAs in every part might not yield accurate results. Thus, we give detailed experiments and show that adding the LoRA module to the MLP emotion encoder $\mathbf{E_{emotion}}$ and MLP motion decoder $\mathbf{D_{motion}}$ performs best. It might be because the attention module in the transformer decoder $\mathbf{D}$ has learned more general audio-to-3D knowledge, which has little direct impact on emotions. In contrast, the MLP layers used for encoding emotional information and mapping emotional facial motions have a stronger influence on the output distribution of specific emotions and are therefore more closely tied to emotional knowledge. In this stage, the emotion encoder $\mathbf{E_{emotion}}$ takes emotion information $\mathbf{e_n}$ as input to produce emotion embedding. This emotion embedding is then added to the facial motion features and fed into the transformer decoder $\mathbf{D}$ to obtain the predicted emotional facial motion features.

Specifically, for each emotion $\mathbf{e_i}$, we learn a distinct emotion adapter $\mathbf{A^i_{emotion}}$ and $\mathbf{A^i_{motion}}$ for the emotion encoder $\mathbf{E_{emotion}}$ and the motion decoder $\mathbf{D_{motion}}$ respectively. When considering the emotion encoder $\mathbf{E_{emotion}}$, to assimilate the emotional knowledge of $\mathbf{e_i}$, it is sufficient to train the parameters $\Delta W^i_{emotion}$ within the emotion adapter $\mathbf{A^i_{emotion}}$ while keeping the parameters $W_{emotion}$ of the emotion encoder $\mathbf{E_{emotion}}$ fixed. During the inference phase, by incorporating $\Delta W^i_{emotion}$ and $W_{emotion}$ as indicated in (1), we can generate an enhanced emotion encoder $\mathbf{E^i_{emotion}}$ that embodies the knowledge of the specific emotion $\mathbf{e_i}$. An analogous approach is applied to the motion decoder $\mathbf{D_{motion}}$ to learn emotional knowledge. After the above steps, we obtain the $\mathbf{E^i_{emotion}}$ and $\mathbf{D^i_{motion}}$ for each emotion $\mathbf{e_i}$. In this way, this allows us to get the emotional 3D mesh as in (4):

$$\hat{\mathbf{y}}_\mathbf{t} = \mathbf{D^i_{motion}}(\mathbf{D}(\mathbf{E_{audio}}(\mathbf{a}), \mathbf{E^i_{emotion}}(\mathbf{e_i}) + \mathbf{E_{motion}}(\hat{\mathbf{y}}_{<\mathbf{t}}))), \tag{4}$$

where $\mathbf{i}$ represents a specific category of emotion.

3) Merging Adapters With Frame-Wise Emotional Guidance: After learning the emotion-specific LoRA adapters individually, to facilitate the performance and the ease of usage, we then combine these adapters for emotional talking portrait generation. Directly using current LoRA merge techniques in the LLM (Large Language Model) field [37], [38], [39], [40] might damage each adapter's individual emotional knowledge, especially since we have eight emotion-specific adapters. To address this limitation, as shown in stage 3 of Fig. 3, we first combine and solve the clash in the emotional adapters inspired by TIES [40]. Besides, different from previous works that utilize the video-level annotations for emotion, in this stage, since our LoRA adapters can be considered as a basis for different emotions, we utilize more fine-grained guidance, i.e., frame-wise guidance, as the condition for the emotional talking head generation.

Algorithm 1: TIES-Based Merging of Emotion-Specific LoRA Adapters With Frame-Wise Emotional Guidance

Input: Base emotion encoder E_{emotion} with weights W_{emotion}, motion decoder D_{motion} with weights W_{motion}, trained emotion-specific LoRA adapters $\{A^i_{\text{emotion}}, A^i_{\text{motion}}\}^n_{i=1}$ with $\{\Delta W^i_{\text{emotion}}, \Delta W^i_{\text{emotion}}\}^n_{i=1}$, pretrained POSTER-V2, training videos with frames $\{\text{frame}_t\}$, audio a, trimming threshold $\tau\%$.

Output: Merged emotional LoRA adapters $\{A^{\text{merge}}_{\text{emotion}}, A^{\text{merge}}_{\text{motion}}\}$ with $\{\Delta W^{\text{merge}}_{\text{emotion}}, \Delta W^{\text{merge}}_{\text{motion}}\}$.

Frame-wise emotional guidance extraction
for each training sequence and each frame t **do**
  $e_{\text{frame},t} \leftarrow \text{POSTER-V2}(\text{frame}_t) \in \mathbb{R}^n$ ▷n-dimensional fine-grained multi-label emotional features
end for
TIES merging for each module $m \in \{\textbf{emotion}, \textbf{motion}\}$
for $m \in \{\text{emotion}, \text{motion}\}$ **do**
  $\Delta W^{\text{merge}}_{\text{m}} \leftarrow \text{TIES-Merge}(\{\Delta W^i_m\}^n_{i=1}, \tau\%)$
end for
$E^{\text{merge}}_{\text{emotion}} \leftarrow$ base emotion encoder with weights $W_{\text{emotion}} + \Delta W^{\text{merge}}_{\text{emotion}}$
$D^{\text{merge}}_{\text{motion}} \leftarrow$ base motion decoder with weights $W_{\text{motion}} + \Delta W^{\text{merge}}_{\text{motion}}$
Fine-tuning merged adapters with frame-wise emotional guidance
Freeze $W_{\text{emotion}}, W_{\text{motion}}$; update only $\Delta W^{\text{merge}}_{\text{emotion}}, \Delta W^{\text{merge}}_{\text{motion}}$
for each training sequence with audio a **do**
  $\hat{y}_t \leftarrow D^{\text{merge}}_{\text{motion}}(D(E_{\text{audio}}(a), E^{\text{merge}}_{\text{emotion}}(e_{\text{frame}}) + E_{\text{motion}}(\hat{y}_{<t})))$
  Compute loss $L \leftarrow loss(\hat{y}_t, y_t)$
Update $\Delta W^{\text{merge}}_{\text{emotion}}, \Delta W^{\text{merge}}_{\text{motion}}$ by minimizing L
end for
return $\{A^{\text{merge}}_{\text{emotion}}, A^{\text{merge}}_{\text{motion}}\}$ with $\{\Delta W^{\text{merge}}_{\text{emotion}}, \Delta W^{\text{merge}}_{\text{motion}}\}$

The specific details of merging adapters are shown in Algorithm 1. We first employ POSTER-V2 [44], a pre-trained face emotion recognition model, to extract the fine-grained, multi-label encoding of the expression information in each video frame. This model generates an 8-dimensional emotion vector for each frame, where each dimension represents a distinct facial expression and its corresponding intensity. A visual comparison between video-level labels and frame-wise guidance is presented in Fig. 4. The comparison illustrates that the frame-wise guidance $\mathbf{e_{frame}}$ offers richer detailed expressive features than static video-level labels $\mathbf{e_n}$. Then we combine the trained emotion-specific adapters $\mathbf{A^i}$ to obtain a fine-grained merged emotion adapter $\mathbf{A^{merge}}$:

$$\mathbf{A^{merge}_{emotion}} = \mathbf{M_{TIES}}(\mathbf{A^i_{emotion}}), i = 1, 2, \ldots, n, \tag{5}$$

Authorized licensed use limited to: INSTITUTE OF COMPUTING TECHNOLOGY CAS. Downloaded on October 19,2026 at 08:40:51 UTC from IEEE Xplore. Restrictions apply.

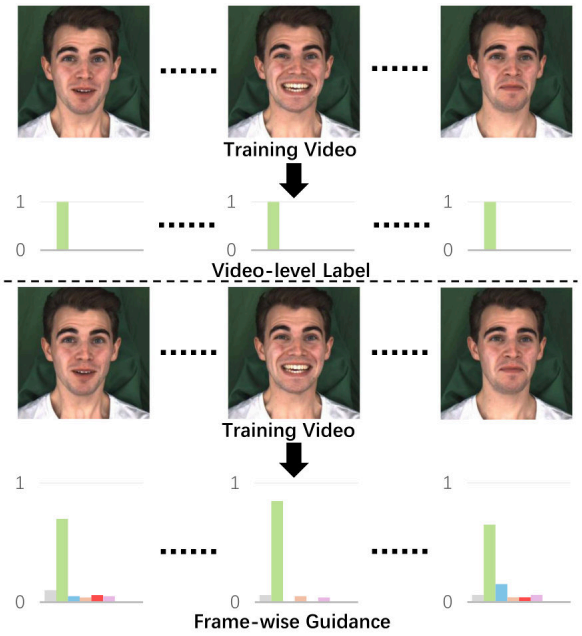

Fig. 4. Visualization of the video-level label and frame-wise guidance for the same training video. The emotion category of this training video is happy. Distinct color bars correspond to various emotions, with the bar height indicating expression intensity. For a given training video, the video-level label is static across all frames; however, the actual expressions vary dynamically from frame to frame. The frame-level guidance enables a more granular capture of the expressive features inherent in individual frames of the video.

$$\mathbf{A}^{\mathbf{merge}}_{\mathbf{motion}} = \mathbf{M}_{\mathbf{TIES}}(\mathbf{A}^{\mathbf{i}}_{\mathbf{motion}}), i = 1, 2, \ldots, n, \quad (6)$$

where $\mathbf{n}$ represents the total number of emotion types. Equations (5) and (6) both execute the same operation, $\mathbf{M}_{\mathbf{TIES}}$, whose details are provided in Algorithm 2. Specifically, in $\mathbf{M}_{\mathbf{TIES}}$ process, we first trim parameters ΔW^i within each emotion-specific adapter that exhibits low-magnitude changes compared to the base model parameters, ensuring only the top-$\tau\%$ parameter values are retained. We then resolve parameter sign conflicts between each emotion-specific adapter. The final step is to merge only those parameters in each adapter that align with the collectively determined sign. For instance, as depicted in stage 3 of Fig. 3, (5) describes $\mathbf{M}_{\mathbf{TIES}}$ process where the parameter $\Delta W^i_{emotion}$ of every emotional adapter $\mathbf{A}^{\mathbf{i}}_{\mathbf{emotion}}$ undergoes trim and merge operation to produce a unified parameter $\Delta W^{merge}_{emotion}$ of merged adapter $\mathbf{A}^{\mathbf{merge}}_{\mathbf{emotion}}$. This combined $\Delta W^{merge}_{emotion}$ is then integrated with the parameter $W_{emotion}$ of emotion encoder $\mathbf{E}_{\mathbf{emotion}}$ to form the parameters of $\mathbf{E}^{\mathbf{merge}}_{\mathbf{emotion}}$, as prescribed by (1). At this frame-wise guidance training stage, we also only need to train the parameters $\Delta W^{merge}_{emotion}$ within the merged emotional adapter $\mathbf{A}^{\mathbf{merge}}_{\mathbf{emotion}}$ and keep the parameter $W_{emotion}$ of emotion encoder $\mathbf{E}_{\mathbf{emotion}}$ fixed. After the same $\mathbf{M}_{\mathbf{TIES}}$ operation, we can get $\mathbf{D}^{\mathbf{merge}}_{\mathbf{motion}}$ from (6).

After obtaining the merged emotional adapters $\mathbf{A}^{\mathbf{merge}}_{\mathbf{emotion}}$ and $\mathbf{A}^{\mathbf{merge}}_{\mathbf{motion}}$, we retrain them on the fine-grained emotional guidance $\mathbf{e}_{\mathbf{frame}}$. This procedure is formulated as:

$$\hat{\mathbf{y}}_{\mathbf{t}} = \mathbf{D}^{\mathbf{merge}}_{\mathbf{motion}}(\mathbf{D}(\mathbf{E}_{\mathbf{audio}}(\mathbf{a}), \mathbf{E}^{\mathbf{merge}}_{\mathbf{emotion}}(\mathbf{e}_{\mathbf{frame}}) + \mathbf{E}_{\mathbf{motion}}(\hat{\mathbf{y}}_{<\mathbf{t}}))). \quad (7)$$

Algorithm 2: TIES-Merge: Trim-Sign-Merge for Adapters

function TIES-Merge$\{\Delta W^i\}_{i=1}^n, \tau\%$
  Trim low-magnitude updates
  for $i = 1$ to n **do**
    $v \leftarrow \text{flatten}(|\Delta W_i|)$
    $N \leftarrow \text{length}(v)$
    $K \leftarrow \lfloor \tau\% \cdot N \rfloor$
    $\theta \leftarrow \text{kth_largest}(v, K)$
    for each parameter index k **do**
      if $|\Delta W^i[k]| < \theta$ **then**
        $\Delta W^i[k] \leftarrow 0$
      end if
    end for
  end for
  Resolve sign conflicts
  for each parameter index k **do**
    $s_k \leftarrow \text{sign}(\sum_{i=1}^n \text{sign}(\Delta W^i[k]))$ ▷Compute consensus sign
    for $i = 1$ to n **do**
      if $\Delta W^i[k] \neq 0$ **and** $\text{sign}(\Delta W^i[k]) \neq s_k$ **then**
        $\Delta W^i[k] \leftarrow 0$
      end if
    end for
  end for
  Merge surviving updates
  for each parameter index k **do**
    $\Delta W^{\text{merge}}[k] \leftarrow \text{Agg}(\{\Delta W^i[k]\}_{i=1}^n)$ ▷Average over non-zeros
  end for
  return ΔW^{merge}
end function

This fusion strategy allows the model to combine each emotion-specific adapter $\mathbf{A}^{\mathbf{i}}$ based on the expression composition of each frame, rather than relying on a static merging technique. As a result, the fused fine-grained emotional adapter $\mathbf{A}^{\mathbf{merge}}$ preserves the emotional knowledge acquired by individual adapters $\mathbf{A}^{\mathbf{i}}$ and attains enhanced precision in emotion control compared to single emotion-specific adapters $\mathbf{A}^{\mathbf{i}}$.

4) Loss Function: We involve several losses to calculate the difference between the predicted 3D mesh $\hat{\mathbf{Y}}_{\mathbf{t}} = (\hat{\mathbf{y}}_{\mathbf{1}}, \ldots, \hat{\mathbf{y}}_{\mathbf{T}})$ and the ground truth $\mathbf{Y}_{\mathbf{t}} = (\mathbf{y}_{\mathbf{1}}, \ldots, \mathbf{y}_{\mathbf{T}})$, where $\mathbf{T}$ is the number of video frames. The first loss is the 3D mesh reconstruction loss L_{rec}, which is a frame-wise Mean Squared Error (MSE) between the predicted 3D face mesh and the ground truth mesh:

$$L_{rec} = \sum_{\mathbf{t}=1}^{\mathbf{T}} \|\hat{\mathbf{y}}_{\mathbf{t}} - \mathbf{y}_{\mathbf{t}}\|^2, \quad (8)$$

this loss is also used for training the basic auido-to-3D mesh model in stage 1 of the Fig. 3.

Besides, for emotion generation, we find single L_{rec} can not capture the detailed face changes in both emotion and lip-synchronization. Thus, we involve the velocity loss L_{vel} for this purpose similar to [43]. In detail, L_{vel} computes the L_1 distance between the consecutive frames between predicted outputs

Authorized licensed use limited to: INSTITUTE OF COMPUTING TECHNOLOGY CAS. Downloaded on October 19,2026 at 08:40:51 UTC from IEEE Xplore. Restrictions apply.

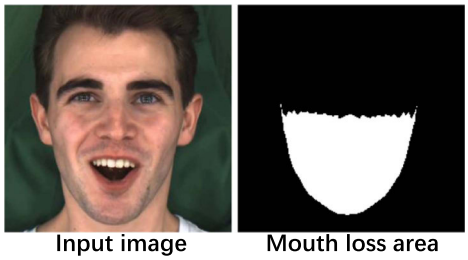

Fig. 5. Visualization of the masked mouth area loss.

and ground truth vertices, which can be expressed as:

$$L_{vel} = \sum_{\mathbf{t=1}}^{\mathbf{T-1}} \left\| (\hat{\mathbf{y}}_{\mathbf{t+1}} - \hat{\mathbf{y}}_{\mathbf{t}}) - (\mathbf{y}_{\mathbf{t+1}} - \mathbf{y}_{\mathbf{t}}) \right\|^2. \tag{9}$$

Furthermore, to better generate a lively expression, we propose a masked loss L_{mou} in the mouth region. It attends more closely to facial vertices in the regions that are crucial for audio-visual synchronization and conveying emotions, such as the lower half of the face. The visualization of the masked mouth loss area is illustrated in Fig. 5. Optimizing this mouth area loss significantly improves both the accuracy of lip movements and the overall expressiveness of the synthesized talking portraits. The mouth area loss can be written as:

$$L_{mou} = \sum_{\mathbf{t=1}}^{\mathbf{T}} \sum_{\mathbf{v=1}}^{\mathbf{V}} \left\| \hat{\mathbf{y}}_{\mathbf{t,v}} - \mathbf{y}_{\mathbf{t,v}} \right\|^2, \tag{10}$$

where $\mathbf{V}$ represents the number of vertices in the mouth loss area.

Overall, the losses for our LoRA adapter training (stage 2 and stage 3) can be expressed as:

$$L_{total} = \lambda_r L_{rec} + \lambda_v L_{vel} + \lambda_m L_{mou}, \tag{11}$$

where $\lambda_r, \lambda_v, \lambda_m$ are the hyper-parameters, respectively.

C. Photo-Realistic Face Render From 3D Mesh

After getting the predicted 3D meshes $\hat{\mathbf{Y}}_\mathbf{t} = (\hat{\mathbf{y}}_\mathbf{1}, \ldots, \hat{\mathbf{y}}_\mathbf{T})$, we render them to rendering images $\mathbf{R_t} = (\mathbf{r_1}, \ldots, \mathbf{r_T})$ by using the shape, texture, light, and camera 3DMM coefficients, where $\mathbf{T}$ is the number of video frames. These synthesized rendering images $\mathbf{R_t} = (\mathbf{r_1}, \ldots, \mathbf{r_T})$ are then transferred into photo-realistic face images $\mathbf{I_t} = (\mathbf{i_1}, \ldots, \mathbf{i_T})$ through our face generator $\mathbf{G}$.

Specifically, we utilize the StyleUNet [45] as our fundamental translation generator. The StyleUNet employs an encoder-decoder architecture specifically designed for image-to-image translation tasks. Analogous to UNet [46], the encoder captures multi-scale features which are then relayed to the decoder through skip connections. This StyleUNet structure is trained adversarially using a StyleGAN discriminator [47]. However, if the synthetic rendering image is directly transferred to the realistic image, it is difficult to explicitly control each emotion because the latent space of StyleUNet is highly coupled [33]. To address this, we incorporated emotion information into the face generator for precise emotion control. Specifically, the emotion label $\mathbf{e_n}$ is mapped to the latent space and then added to the latent. Additionally, to enhance audio-visual synchronization in the generated talking videos, we integrate the audio features into the face generator. These audio features, extracted by the pre-trained speech model wav2vec 2.0 [42] from the input driving audio $\mathbf{a}$, are also integrated into the latent space. This render-to-image process is shown in Fig. 2, which can be written as follows:

$$\mathbf{i_t} = \mathbf{G}_\phi(\mathbf{a}, \mathbf{e_n}, \mathbf{r_t}), \tag{12}$$

where ϕ represents the parameters of face generator $\mathbf{G}$. Regarding loss functions, we adhere to the methodology outlined in [45], employing L1 loss and perceptual loss with a VGG19 network [48] during the training phase. Additionally, we incorporate Generative Adversarial Network (GAN) loss to facilitate adversarial learning.

IV. Experiments and Results

A. Experimental Setup

1) Dataset and Implementation Details: We train and evaluate our framework on the MEAD [9] and CREMA-D [49] dataset. MEAD is a high-quality emotional talking-head video dataset with 8 kinds of emotions. Here we select 4 subjects, two males and two females, following the same test cases mentioned in EAT [17] and EDTalk [18]. For CREMA-D, we select 2 subjects as test cases. The training videos are sampled to 25 FPS and the audio sample rate is 16KHz. Since the MEAD dataset does not come with 3D meshes, for each video, we first perform parametric model fitting to extract 3DMM coefficients, and then input these 3DMM coefficients into FLAME [50] to get 3D meshes as training data to train our audio-to-3D mesh model. Our method is implemented in PyTorch and trained on a single NVIDIA 4090 GPU. Using the MEAD dataset as an example, we train the audio-to-3D mesh module with eight emotions, each with approximately 100 seconds of video. This stage uses the Adam optimizer with a learning rate of $1e-4$ and a batch size of 1. Each emotion-specific LoRA adapter is trained for 0.5 hours on the corresponding emotion videos; the hyper-parameters λ_r, λ_v, and λ_m in this stage are set to 1, 6, and 5, respectively. In the merging adapters training stage, we fine-tune the merged adapter for 6 hours. The λ_r, λ_v, and λ_m are set to 1, 1, and 3, respectively, and the trimming threshold $\tau\%$ is set to 50%. Finally, the face renderer is trained for 14 hours using Adam with a learning rate of 2e-3 and a batch size of 4.

2) Evaluation Metrics: We consider the following four types of metrics to measure the generated emotional talking portrait videos:

Emotional accuracy: The emotional accuracy of the talking videos is assessed using the Emotion-Fan classifier [51], as referenced in EAT [17].

Audio-visual synchronization: We use the lip-sync metric Sync-C from SyncNet [32] mentioned in Wav2Lip [1] to evaluate the accuracy of the synchronization between the audio and the produced video. Additionally, the Landmark Velocity Difference (LVD) metric is applied to the mouth (M-LVD) area

Authorized licensed use limited to: INSTITUTE OF COMPUTING TECHNOLOGY CAS. Downloaded on October 19,2026 at 08:40:51 UTC from IEEE Xplore. Restrictions apply.

TABLE I
QUANTITATIVE COMPARISONS WITH THE STATE-OF-THE-ART METHODS ON MEAD [9] AND CREMA-D [49] TEST SET. THE N./A. IN THE ADAPTER PARAMS COLUMN MEANS THAT THESE METHODS TRAIN THE ENTIRE MODEL FOR EMOTIONAL TALKING HEAD GENERATION. TIME DENOTES THE AVERAGE TIME COST OF GENERATING ONE SECOND OF 25 FPS VIDEO. - MEANS THE MODEL ON THE CORRESPONDING DATASET IS NOT AVAILABLE.

	MEAD [9] / CREMA-D [49]								
	Adapter Params	Acc_{emo} ↑	FID ↓	M-LVD ↓	Sync-C ↑	Time	Emotion Control	Type	Modeling
MakeItTalk [20]	N./A.	12.88 / 16.67	85.97 / 51.78	1.79 / 1.76	4.90 / 3.91	2.77s	N./A.	One-shot	2D-based
SadTalker [2]	N./A.	12.88 / 16.67	84.78 / 63.27	1.84 / 1.76	6.93 / **6.85**	4.70s	N./A.	One-shot	2D-based
Aniportrait [22]	N./A.	12.88 / 16.67	42.45 / 37.87	1.86 / 1.88	4.82 / 4.37	32.79s	N./A.	One-shot	2D-based
Hallo [6]	N./A.	12.88 / 16.67	41.19 / 32.45	1.76 / 1.83	6.29 / 6.07	53.33s	N./A.	One-shot	2D-based
Echomimic [8]	N./A.	12.88 / 16.67	38.98 / 33.00	1.74 / 1.80	5.89 / 5.06	35.57s	N./A.	One-shot	2D-based
EAMM [13]	N./A.	13.49 / 29.17	83.92 / 63.32	1.77 / 1.79	3.51 / 4.32	3.54s	Reference	One-shot	2D-based
PD-FGC [14]	N./A.	42.94 / 54.17	80.21 / 42.49	1.88 / 1.89	5.39 / 4.00	1.86s	Reference	One-shot	2D-based
EDTalk-V [18]	N./A.	54.75 / 45.83	49.31 / 38.77	1.87 / 1.78	**7.97** / 6.69	4.77s	Reference	One-shot	2D-based
SDTFG [11]	N./A.	- / 60.42	- / 35.12	- / 1.89	- / 5.23	**0.85s**	Label	One-shot	2D-based
EVP [10]	N./A.	49.25 / -	26.75 / -	1.69 / -	3.82 / -	8.96s	Label	Person-specific	3D-based
EAT [17]	13.9M	77.30 / 58.33	47.53 / 36.98	1.73 / 1.87	7.43 / 6.28	4.98s	Label	One-shot	2D-based
Ours	**0.2M**	**83.89 / 66.67**	**21.79 / 20.52**	**1.67 / 1.70**	5.03 / 4.96	5.01s	Fine-grained Label	Person-specific	3D-based

to evaluate the accuracy of lip movements in the synthesized videos, following the EVP [10].

Visual quality: The Frechet Inception Distance (FID) score [52] is utilized to measure the image quality of synthesized emotional talking videos.

B. Comparison With Other Methods

We conduct quantitative comparisons with other state-of-the-art methods on the MEAD [9] test dataset, including MakeItTalk [20], SadTalker [2], Aniportrait [22], Hallo [6], Echomimic [8], SDTFG [11], EVP [10], EAMM [13], PD-FGC [14], EAT [17], and EDTalk [18]. The results of EVP [10] are tested on subjects provided by the author. For EAMM [13], PD-FGC [14], and EDTalk [18], which derive expressions from emotional videos, we selected videos with corresponding expressions from the MEAD dataset as emotion references. Table I shows our method outperforms others in terms of visual quality and emotional accuracy, demonstrating its proficiency in learning precise emotion features through fine-grained frame-wise expression guidance. Regarding efficiency, methods such as EVP [10], EAMM [13], PD-FGC [14], MakeItTalk [20], SadTalker [2], EDTalk [18], Aniportrait [22], Hallo [6], Echomimic [8], which train the entire network for talking head generation, are found to be parameter-inefficient. Although EAT [17] employs a fine-tuning approach, its emotional adapter comprises 13.9 M parameters. In contrast, our emotional adapter contains a mere 0.2 M parameters, showing the superior efficiency of our emotion control method. Furthermore, we report the average inference time required to generate one second of video at 25 fps. Diffusion-based methods, such as Aniportrait [22], Hallo [6], and Echomimic [8], require substantially more time for inference. In contrast, our approach and other GAN-based methods produce results within seconds. Among 3D-based approaches, our method achieves faster inference than EVP [10]. For audio-visual synchronization, our method exhibits performance comparable to other methods. The superior Sync-C scores may not necessarily ensure better results, as the SyncNet [32] model is exclusively trained on neutral emotion videos. And our method achieves superior M-LVD scores, indicating that the lip movements generated by our model more closely approximate the ground truth.

We further make visual comparisons with other state-of-the-art methods. As shown in Figs. 6 and 7, our method shows the results with more vivid and faithful facial emotional details that closely resemble the original expression of the ground truth, particularly in the lip corners, wrinkles, eyes, and eyebrow areas. Moreover, the talking head videos generated by our method boast higher definition, showing clear teeth and skin details compared to other methods. Furthermore, our approach stands out in identity preservation, which can produce more personalized speaking and emotional styles closer to ground truth identity. The emotions generated by PD-FGC [14] and EAMM [13] exhibit limited consistency with the designated emotion labels, and they often result in blurred facial details. While EAT [17] and EDTalk [18] are capable of producing expressions that align with emotion labels, they struggle to preserve personalized speech and emotional styles that are consistent with ground truth identity.

C. User Study

To further evaluate the quality of the generated emotional talking videos, we conduct a user study with 14 participants. This study assesses the talking videos across three dimensions: audio-visual synchronization, video quality, and emotional accuracy. Scores range from a minimum of 1 to a maximum of 5. We selecte 12 audio clips from the MEAD test dataset, covering 8 emotions, to generate talking video samples for different methods: EVP [10], PD-FGC [14], SadTalker [2], EAT [17], EDTalk [18], and our method. The user study results are shown in Table II. Our method achieves the best scores for video quality and emotion accuracy, while demonstrating comparable performance in audio-visual synchronization.

D. Ablation Studies

We perform detailed ablation studies to show the effectiveness of each design in our methods.

Authorized licensed use limited to: INSTITUTE OF COMPUTING TECHNOLOGY CAS. Downloaded on October 19,2026 at 08:40:51 UTC from IEEE Xplore. Restrictions apply.

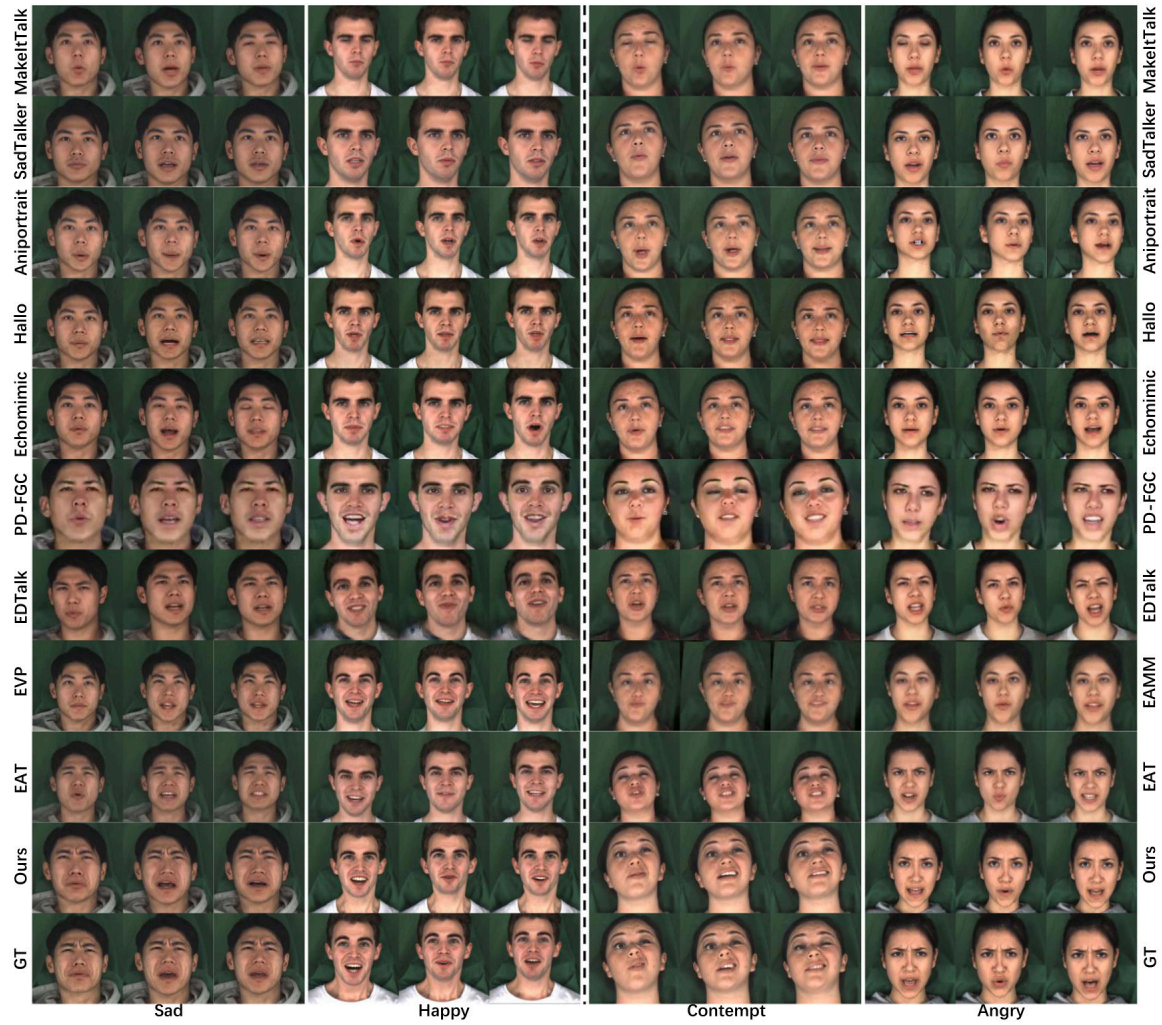

Fig. 6. Visual comparison with other state-of-the-art methods for emotional talking portrait video on the MEAD [9] test dataset. This figure primarily illustrates the comparative outcomes of our method relative to MakeItTalk [20], SadTalker [2], PD-FGC [14], EVP [10], EAMM [13], EAT [17], EDTalk [18], Aniportrait [22], Hallo [6], and Echomimic [8] with respect to emotional accuracy, visual quality, identity preservation, and lip synchronization. For a more detailed examination, kindly enlarge the image or view our supplemental video.

TABLE II
USER STUDY ON MEAD TEST DATASET

	SadTalker	PD-FGC	EDTalk	EVP	EAT	Ours	GT
Lip-sync	3.09	3.30	**3.53**	2.61	3.40	3.43	4.79
Video quality	3.97	1.91	3.09	3.34	3.30	**3.99**	4.64
Acc_{emo}	1.61	2.56	3.52	1.84	3.66	**3.68**	4.68

TABLE III
QUANTITATIVE RESULTS OF THE ABLATION STUDY FOR LORA ADAPTER ON DIFFERENT MODULES AND MASKED MOUTH LOSS

	Acc_{emo} ↑	FID ↓	M-LVD ↓	Sync-C ↑
LoRA on attention	85.42%	26.37	1.66	5.26
w/o mouth loss	83.33%	23.87	1.65	5.48
Ours	**89.58%**	**19.67**	**1.61**	**5.85**

1) LoRA Adapter on Different Modules: We conduct the ablation experiment to confirm the efficacy of applying LoRA in our adaptable MLP emotion encoder $\mathbf{E_{emotion}}$ and motion decoder $\mathbf{D_{motion}}$ to learn expression-related knowledge. We explore different adapter placements, including the attention module in the transformer decoder $\mathbf{D}$ and the MLP emotion encoder $\mathbf{E_{emotion}}$ and motion decoder $\mathbf{D_{motion}}$. The visual analysis of Fig. 8 demonstrates that applying the LoRA to the attention module within the transformer decoder $\mathbf{D}$ results in a mouth area with a reduced opening, more closely resembles the mouth movements associated with neutral expressions. In contrast, employing the LoRA on both the MLP emotion encoder $\mathbf{E_{emotion}}$ and motion decoder $\mathbf{D_{motion}}$ produces a mouth that displays characteristics of surprise, aligning more closely with the ground truth. The quantitative results presented in Table III reveal that applying LoRA to the attention module of the transformer decoder $\mathbf{D}$ not only fails to learn the more precise expression knowledge but also damages generic audio-visual synchronization. These findings indicate that the MLP emotion encoder $\mathbf{E_{emotion}}$ and motion decoder $\mathbf{D_{motion}}$ modules have a stronger correlation with expressions, while the attention

Authorized licensed use limited to: INSTITUTE OF COMPUTING TECHNOLOGY CAS. Downloaded on October 19,2026 at 08:40:51 UTC from IEEE Xplore. Restrictions apply.

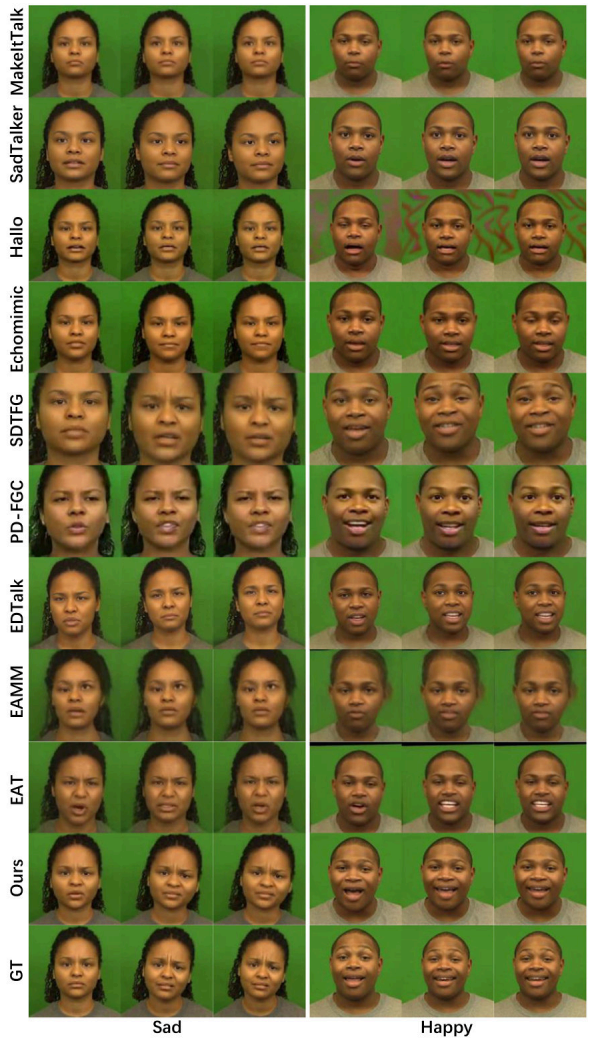

Fig. 7. Visual comparison with other state-of-the-art methods for emotional talking portrait video on the CREMA-D [49] test dataset. This figure primarily illustrates the comparative outcomes of our method relative to MakeItTalk [20], SadTalker [2], SDTFG [11], PD-FGC [14], EAMM [13], EAT [17], EDTalk [18], Hallo [6], and Echomimic [8] with respect to emotional accuracy, visual quality, identity preservation, and lip synchronization.

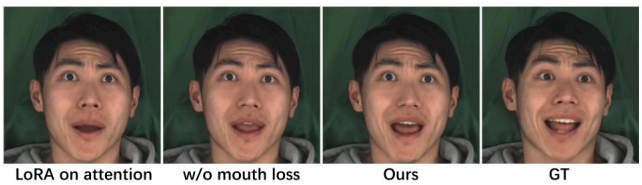

Fig. 8. Qualitative results of ablation study for LoRA adapter on different modules and masked mouth area loss. The emotion of the talking videos is surprising. The expression we generated is closer to ground truth, with a more widely opened mouth, reflecting more obvious surprised characteristics.

module in the transformer decoder $\mathbf{D}$ has learned more general audio-to-3D mesh knowledge that should remain largely unchanged. By strategically applying the LoRA to the MLP emotion encoder $\mathbf{E_{emotion}}$ and motion decoder $\mathbf{D_{motion}}$, we achieve efficient and effective emotional transfer, while preserving the general audio-to-3D capabilities of the base model.

Fig. 9. Qualitative results of the ablation study for different training methods. The emotion of the talking videos is happy. The synthesized mouth exhibits a wider opening and more closely matches the ground truth.

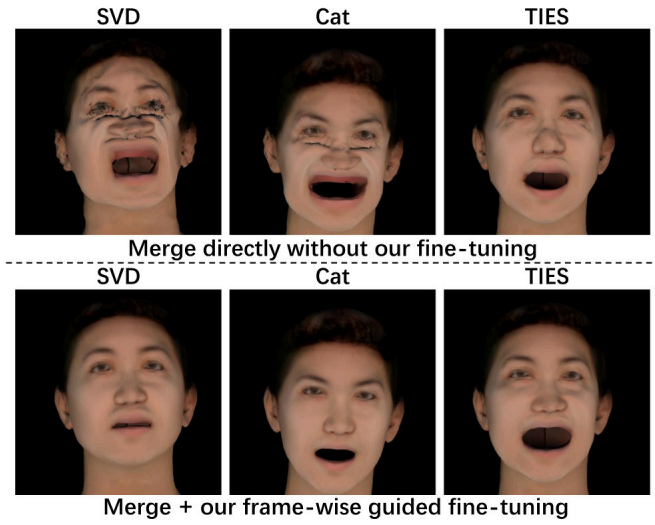

Fig. 10. Qualitative results of ablation study for our adapter merge via frame-wise emotion-guided fine-tuning. Our fine-tuning with frame-wise emotional guidance can eliminate artifacts in the generated 3D mesh.

2) Mouth Area Loss: We perform an ablation study to assess the impact of masked mouth area loss. Results presented in Table III and Fig. 8 indicate that mouth area loss enhances the vividness and accuracy of generated facial expressions, in addition to improving audio-visual synchronization and visual quality.

3) Adapter Merge Via Frame-Wise Emotional Guidance: Current adapter merging techniques, such as Singular Value Decomposition (SVD) [37], concatenation (Cat) [37], TIES [40], if used directly after merging, tend to destroy the individual knowledge captured by each adapter, especially when dealing with a large number of adapters. As shown in the upper part of Fig. 10, directly employing SVD, Cat, and TIES to amalgamate individual emotional adapters results in generated emotional 3D meshes with pronounced artifacts. This suggests that such straightforward merging techniques can compromise the expression knowledge acquired by each emotional adapter. However, following our finetune with frame-wise emotional guidance, the merged adapter demonstrates an enhanced ability to accurately reconstruct emotional knowledge.

Authorized licensed use limited to: INSTITUTE OF COMPUTING TECHNOLOGY CAS. Downloaded on October 19,2026 at 08:40:51 UTC from IEEE Xplore. Restrictions apply.

TABLE IV
QUANTITATIVE RESULTS OF THE LORA COMBINATION IN AUDIO-TO-3D MESH GENERATION

	Adapter Params	Acc_{emo} ↑	FID ↓	M-LVD ↓
Full finetune	N.A.	97.92%	25.92	1.62
Each adapters before merge	1.94M	89.58%	19.67	1.61
SVD + frame-wise guidance finetune	**0.24M**	**100.00**%	21.78	1.61
Cat + frame-wise guidance finetune	1.94M	**100.00**%	22.39	**1.59**
Video-level label finetune	**0.24M**	89.58%	19.09	1.60
Ours	**0.24M**	**100.00**%	**18.08**	**1.59**

Fig. 11. Qualitative results of the ablation study for our audio-guided face render. The first image on the left is the rendered image used as input for the face render. The emotion category of the talking videos is happy.

Subsequently, we conduct ablation studies for different audio-to-3D mesh training methods. Our approach, which utilizes frame-wise emotional guidance to guide TIES merging, is compared to several alternative methods: full model fine-tuning, individual emotion-specific adapter learning via LoRA, frame-wise guidance guided SVD and Cat merging, and video-level label guided TIES merging. As shown in Fig. 9, our methods yield a mouth that more closely matches the ground truth. Moreover, as shown in Table IV, after our frame-wise emotion guidance finetune, these merging methods outperform both individual emotion-specific adapters and the fully fine-tuned model in emotional accuracy, audio-visual synchronization, and visual quality. Most importantly, our approach requires training only an emotional adapter with 0.24 M parameters. Furthermore, a solitary adapter suffices to regulate all emotions, offering convenience significantly superior to the need for individual adapters for each emotion. Moreover, as presented in Table IV and Fig. 9, our approach leverages frame-level guidance during fine-tuning, yielding more precise control over emotional expression than fine-tuning with video-level labels. This demonstrates that our method offers effective and efficient emotional transfer by using frame-wise feature guided TIES merge.

4) Audio-Guided Face Render: We conduct the ablation study on the audio-guided face render. Our results, presented in Fig. 11 and supplement video, show that inputting the features of driven audio into the face render can enhance the audio-visual synchronization and naturalness of the generated emotional talking head video.

E. Generalizability Analysis

Our method is a person-specific approach for generating high-quality emotional talking head videos, which is also practical since our method can produce more personalized speaking and emotional styles. Compared to single-image animation methods, our approach excels in the quality of appearance details and identity preservation. Moreover, one-shot methods typically require enormous emotional images for training, while our method trains on about 100 seconds of emotional videos. Despite our method can not directly infer a new identity without specific training, it can be driven by in-the-wild audio as in Fig. 12. And in the supplementary video, we provide talking portrait videos driven by in-the-wild audio for a more intuitive demonstration.

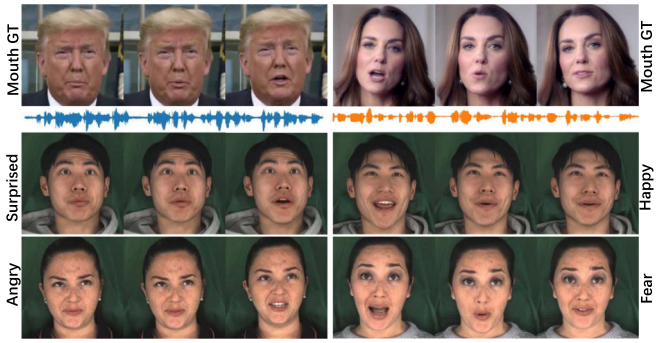

Fig. 12. The emotional portrait results driven by in-the-wild audio. We give the mouth ground truth as a reference. Our method accurately generates emotions while ensuring lip synchronization that aligns with the ground truth mouth.

F. Emotion Editing

Our approach utilizes fine-grained frame-wise emotional guidance, so the learned emotional features are continuous. This enables editing capabilities for the generated talking videos, including modifying the emotion category and intensity level. In addition to supporting the use of a single interpolated emotional label like EVP [10] and EAT [17] to control emotion over the whole video, our method additionally supports a novel emotion editing technique - dynamically modifying the emotions frame by frame during the video. These manipulation capabilities are visually illustrated in Fig. 1 and supplementary.

G. Limitation and Future Work

Although our method achieves SOTA performance on emotional talking head generation, we still have limitations. The generated video will show a slight flicker when the emotion changes during a video. This limitation arises from the challenge of ensuring completely consistent environmental settings when collecting training videos with different emotions, necessitating uniformity in aspects such as character hairstyles, clothing, and lighting conditions. Consequently, our face render is unable to perfectly translate the continuous emotional representation from the 3D mesh to the 2D image. We give the examples in our supplemental video.

Among various control modalities, our method focuses on expression control in talking-head generation. For head pose control of the generated portrait, we extract FLAME head pose parameters from a provided reference video to guide the generation. The gaze of the generated portrait is directly driven by the audio signal, and we do not explicitly model or control eye gaze direction. In future work, we will explore finer-grained gaze control.

Authorized licensed use limited to: INSTITUTE OF COMPUTING TECHNOLOGY CAS. Downloaded on October 19,2026 at 08:40:51 UTC from IEEE Xplore. Restrictions apply.

V. Conclusion

In this paper, we present a novel and efficient method for audio-driven emotional talking portrait generation by considering the scarcity of emotional talking head video datasets and the accuracy of emotional labels. Thus, we propose a multi-stage method to solve these problems. Firstly, we train the basic audio-to-3D mesh model on the neutral expression dataset. Subsequently, we employ LoRA to train emotion-specific adapters on their respective emotional videos for efficient emotional transfer. After that, these adapters are merged, and frame-level fine-grained emotional guidance is utilized to further refine the training of the merged adapter, thereby achieving more precise emotional control. Finally, the produced 3D emotional meshes are input into our audio-guided face renderer for 3D-to-2D video conversion, resulting in high-quality and faithful emotional talking portrait videos. Following our pipeline, our method achieves state-of-the-art performance on the emotional talking portrait generation.

References

[1] 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 *Proc. ACM Int. Conf. Multimedia*, 2020, pp. 484–492.

[2] W. Zhang et al., "Sadtalker: Learning realistic 3D motion coefficients for stylized audio-driven single image talking face animation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2023, pp. 8652–8661.

[3] Z. Ye et al., "Audio-driven talking face video generation with dynamic convolution kernels," *IEEE Trans. Multimedia*, vol. 25, pp. 2033–2046, 2023.

[4] X. Wang, Q. Xie, J. Zhu, L. Xie, and O. Scharenborg, "AnyoneNet: Synchronized speech and talking head generation for arbitrary persons," *IEEE Trans. Multimedia*, vol. 25, pp. 6717–6728, 2023.

[5] S. Shen et al., "SD-NeRF: Towards lifelike talking head animation via spatially-adaptive dual-driven NeRFs," *IEEE Trans. Multimedia*, vol. 26, pp. 3221–3234, 2024.

[6] M. Xu et al., "Hallo: Hierarchical audio-driven visual synthesis for portrait image animation," 2024, *arXiv:2406.08801*.

[7] Y. Ma et al., "TalkCLIP: Talking head generation with text-guided expressive speaking styles," *IEEE Trans. Multimedia*, vol. 27, pp. 6335-6346, 2025.

[8] Z. Chen, J. Cao, Z. Chen, Y. Li, and C. Ma, "Echomimic: Lifelike audio-driven portrait animations through editable landmark conditions," in *Proc. AAAI Conf. Artif. Intell.*, 2025, vol. 39, no. 3, pp. 2403–2410.

[9] K. Wang et al., "Mead: A large-scale audio-visual dataset for emotional talking-face generation," in *Proc. Eur. Conf. Comput. Vis.*, Aug. 2020, pp. 700–717.

[10] X. Ji et al., "Audio-driven emotional video portraits," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 14080–14089.

[11] S. E. Eskimez, Y. Zhang, and Z. Duan, "Speech driven talking face generation from a single image and an emotion condition," *IEEE Trans. Multimedia*, vol. 24, pp. 3480–3490, 2022.

[12] S. Sinha, S. Biswas, R. Yadav, and B. Bhowmick, "Emotion-controllable generalized talking face generation," in *Proc. Int. Joint Conf. Artif. Intell.*, 2022, pp. 1320–1327.

[13] X. Ji et al., "EAMM: One-shot emotional talking face via audio-based emotion-aware motion model," in *Proc. ACM SIGGRAPH Conf. Proc.*, 2022, pp. 1–10.

[14] D. Wang, Y. Deng, Z. Yin, H.-Y. Shum, and B. Wang, "Progressive disentangled representation learning for fine-grained controllable talking head synthesis," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 17979–17989.

[15] S. Gururani, A. Mallya, T.-C. Wang, R. Valle, and M.-Y. Liu, "Space: Speech-driven portrait animation with controllable expression," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2023, pp. 20914–20923.

[16] S. Tan, B. Ji, and Y. Pan, "EMMN: Emotional motion memory network for audio-driven emotional talking face generation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 22146–22156.

[17] Y. Gan, Z. Yang, X. Yue, L. Sun, and Y. Yang, "Efficient emotional adaptation for audio-driven talking-head generation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2023, pp. 22634–22645.

[18] S. Tan, B. Ji, M. Bi, and Y. Pan, "EDTalk: Efficient disentanglement for emotional talking head synthesis," in *Proc. Eur. Conf. Comput. Vis.*, 2025, pp. 398–416.

[19] E. J. Hu et al., "LoRA: Low-rank adaptation of large language models," in *Proc. Int. Conf. Learn. Representations*, 2022.

[20] Y. Zhou et al., "Makelttalk: Speaker-aware talking-head animation," *ACM Trans. Graph.*, vol. 39, no. 6, pp. 1–15, 2020.

[21] Y. Ma et al., "Styletalk: One-shot talking head generation with controllable speaking styles," in *Proc. AAAI Conf. Artif. Intell.*, 2023, pp. 1896–1904.

[22] H. Wei, Z. Yang, and Z. Wang, "Aniportrait: Audio-driven synthesis of photorealistic portrait animation," 2024, *arXiv:2403.17694*.

[23] W. Song et al., "Talkingstyle: Personalized speech-driven 3D facial animation with style preservation," *IEEE Trans. Vis. Comput. Graph.*, vol. 31, no. 9, pp. 4682–4694, Sep. 2025.

[24] B. Dong, B.-Y. Ma, and L. Zhang, "Spatially and temporally optimized audio-driven talking face generation," *Comput. Graph. Forum*, vol. 43, no. 7, 2024, Art. no. e15228.

[25] B. Ma, Y. Cao, and L. Zhang, "Decoupled two-stage talking head generation via Gaussian-landmark-based neural radiance fields," *Comput. Vis. Media*, vol. 11, no. 4, pp. 799–816, 2025.

[26] B. Dong and L. Zhang, "Talking head generation via viewpoint and lighting simulation based on global representation," in *Proc. 33rd ACM Int. Conf. Multimedia*, New York, NY, USA, 2025, pp. 10258–10267.

[27] H. Zhou et al., "Pose-controllable talking face generation by implicitly modularized audio-visual representation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 4176–4186bibpunc.

[28] S. Wang, L. Li, Y. Ding, and X. Yu, "One-shot talking face generation from single-speaker audio-visual correlation learning," in *Proc. AAAI Conf. Artif. Intell.*, 2022, vol. 36, no. 3, pp. 2531–2539.

[29] W. Zhong et al., "Identity-preserving talking face generation with landmark and appearance priors," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2023, pp. 9729–9738.

[30] S.-Y. Chen et al., "Deepfacereshaping: Interactive deep face reshaping via landmark manipulation," *Comput. Vis. Media*, vol. 10, no. 5, pp. 949–963, 2024.

[31] L. Song, W. Wu, C. Qian, R. He, and C. C. Loy, "Everybody's talkin': Let me talk as you want," *IEEE Trans. Inf. Forensics Secur.*, vol. 17, pp. 585–598, 2022.

[32] J. S. Chung and A. Zisserman, "Out of time: Automated lip sync in the wild," in *Proc. Asian Conf. Comput. Vis.*, 2017, pp. 251–263.

[33] Y. Xia, L. Wang, X. Deng, X. Luo, and Y. Liu, "Gmtalker: Gaussian mixture-based audio-driven emotional talking video portraits," 2024, *arXiv:2312.07669v2*.

[34] Q. He et al., "EmoTalk3D: High-fidelity free-view synthesis of emotional 3D talking head," in *Proc. Eur. Conf. Comput. Vis.*, 2024, pp. 55–72.

[35] Z. Peng et al., "EmoTalk: Speech-driven emotional disentanglement for 3D face animation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2023, pp. 20687–20697.

[36] R. Daněček et al., "Emotional speech-driven animation with content-emotion disentanglement," in *Proc. SIGGRAPH Asia Conf. Papers*, 2023, pp. 1–13.

[37] S. Mangrulkar et al., "Peft: State-of-the-art parameter-efficient fine-tuning methods," 2022. [Online]. Available: https://github.com/huggingface/peft

[38] G. Ilharco et al., "Editing models with task arithmetic," in *Proc. Int. Conf. Learn. Representations*, 2023.

[39] L. Yu, B. Yu, H. Yu, F. Huang, and Y. Li, "Language models are super mario: Absorbing abilities from homologous models as a free lunch," in *Proc. Int. Conf. Mach. Learn.*, 2024, pp. 57755–57775.

[40] P. Yadav, D. Tam, L. Choshen, C. A. Raffel, and M. Bansal, "Ties-merging: Resolving interference when merging models," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 36, 2024, pp. 7093–7115.

[41] Y. Fan, Z. Lin, J. Saito, W. Wang, and T. Komura, "Faceformer: Speech-driven 3D facial animation with transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 18770–18780.

[42] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, vol. 33, pp. 12449–12460.

[43] D. Cudeiro, T. Bolkart, C. Laidlaw, A. Ranjan, and M. Black, "Capture, learning, and synthesis of 3D speaking styles," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 10101–10111.

[44] J. Mao et al., "Poster v2: A simpler and stronger facial expression recognition network," *Pattern Recognit.*, vol. 157, 2025, Art. no. 110951.

Authorized licensed use limited to: INSTITUTE OF COMPUTING TECHNOLOGY CAS. Downloaded on October 19,2026 at 08:40:51 UTC from IEEE Xplore. Restrictions apply.

[45] L. Wang et al., "Styleavatar: Real-time photo-realistic portrait avatar from a single video," in *Proc. ACM SIGGRAPH Conf. Proc.*, 2023, pp. 1–10.
[46] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. Med. Image Comput. Comput. Assist. Interv.*, 2015, pp. 234–241.
[47] T. Karras et al., "Analyzing and improving the image quality of stylegan," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 8110–8119.
[48] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 694–711.
[49] H. Cao et al., "CREMA-D: Crowd-sourced emotional multimodal actors dataset," *IEEE Trans. Affect. Comput.*, vol. 5, no. 4, pp. 377–390, Oct.–Dec. 2014.
[50] T. Li, T. Bolkart, M. J. Black, H. Li, and J. Romero, "Learning a model of facial shape and expression from 4D scans.," *ACM Trans. Graph.*, vol. 36, no. 6, pp. 194–1, 2017.
[51] D. Meng, X. Peng, K. Wang, and Y. Qiao, "Frame attention networks for facial expression recognition in videos," in *Proc. IEEE Int. Conf. Image Process.*, 2019, pp. 3866–3870.
[52] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local Nash equilibrium," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 6626–6637.

Yiguo Jiang received the bachelor's and master's degrees in computer science and technology from Jilin University, Changchun, China. He is currently working toward the Ph.D. degree with the Department of Computer and Information Science, University of Macau. His research interests include computer vision and artificial intelligence.

Xiaodong Cun (Member, IEEE) received the B.Sc. degree in computer science from Xidian University, Xi'an, China, and the M.S. and Ph.D. degrees from the Department of Computer and Information Science, University of Macau, Macau, in 2018 and 2021, respectively. He was a senior Researcher with the Visual Computing Center, Tencent AI Lab from 2021 to 2024. He is currently an Assistant Professor with the GVC Lab, Great Bay University, Dongguan, China. His research interests include image/video generation, translation, and editing.

Yong Zhang received the Ph.D. degree in pattern recognition and intelligent systems from the Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2018. From 2015 to 2017, he was a Visiting Scholar with the Rensselaer Polytechnic Institute. He is currently with Meituan, Beijing. His research interests include computer vision and machine learning.

Tao He received the Ph.D. degree from the School of Software, Tsinghua University, Beijing, China, in 2023. He is currently a Postdoctoral Researcher with GRG Banking Equipment Company Ltd., Guangzhou, China. His research interests include the area of large language model and multimodal learning.

Fan Tang received the B.Sc. degree in computer science from North China Electric Power University, Beijing, China, in 2013, and the Ph.D. degree from the Institute of Automation, Chinese Academy of Sciences, Beijing, in 2019. He is currently an Associate Professor with the School of Computer Science and Technology, University of Chinese Academy of Sciences, Beijing. His research interests include computer graphics, computer vision, and machine learning.

Chi-Man Pun received the Ph.D. degree in computer science and engineering from the Chinese University of Hong Kong in 2002, and the M.Sc. and B.Sc. degrees from the University of Macau, Macau. He was the Head of the Department of Computer and Information Science, University of Macau from 2014 to 2019, where he is currently a Professor and in charge of the Image Processing and Pattern Recognition Laboratory. He has authored or coauthored more than 200 refereed papers in many top-tier Journals (including IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY, IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON DEPENDABLE AND SECURE COMPUTING, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, and IEEE TRANSACTIONS ON MULTIMEDIA) and Conferences (including CVPR, ICCV, ECCV, AAAI, NIPS, ICDE, IJCAI, MM, and VR), and co-invented several China/US Patents. He has investigated many externally funded research Projects (including FDCT-NSFC and FDCT-MOST) as PI. He was the recipient of the Macao Science and Technology Award 2014 and Best Paper Award in the 6th Chinese Conference on Pattern Recognition and Computer Vision (PRCV2023). He was the General Chair/Co-Chair/Program Chair for many international conferences such as the 10th and 11th International Conference on Computer Graphics, Imaging and Visualization (CGIV2013, CGIV2014), IEEE International Conference on Visual Communications and Image Processing (VCIP2020), and the International Workshop on Advanced Image Technology (IWAIT2022). He is the Area Chair/SPC/PC for many top-tier conferences such as MM, AAAI, IJCAI, CVPR, ICCV, ECCV, and KDD.He is serving as the Editorial Board member for the journal of Artificial Intelligence (AIJ). He has been listed in the World's Top 2% Scientists by Stanford University since 2020. His research interests include Image Processing and Pattern Recognition; Multimedia and AI Security; Medical Image Analysis.

Authorized licensed use limited to: INSTITUTE OF COMPUTING TECHNOLOGY CAS. Downloaded on October 19,2026 at 08:40:51 UTC from IEEE Xplore. Restrictions apply.