Efficient Emotional Adaptation for Audio-Driven Talking-Head Generation

Yuan Gan\textsuperscript{1,2}, Zongxin Yang\textsuperscript{1,2}, Xihang Yue\textsuperscript{1,2}, Lingyun Sun\textsuperscript{2}, Yi Yang\textsuperscript{1,2}*  
\textsuperscript{1}ReLER, CCAI, Zhejiang University, China  
\textsuperscript{2}College of Computer Science and Technology, Zhejiang University, China

Abstract

Audio-driven talking-head synthesis is a popular research topic for virtual human-related applications. However, the inflexibility and inefficiency of existing methods, which necessitate expensive end-to-end training to transfer emotions from guidance videos to talking-head predictions, are significant limitations. In this work, we propose the Emotional Adaptation for Audio-driven Talking-head (EAT) method, which transforms emotion-agnostic talking-head models into emotion-controllable ones in a cost-effective and efficient manner through parameter-efficient adaptations. Our approach utilizes a pretrained emotion-agnostic talking-head transformer and introduces three lightweight adaptations (the Deep Emotional Prompts, Emotional Deformation Network, and Emotional Adaptation Module) from different perspectives to enable precise and realistic emotion controls. Our experiments demonstrate that our approach achieves state-of-the-art performance on widely-used benchmarks, including LRW and MEAD. Additionally, our parameter-efficient adaptations exhibit remarkable generalization ability, even in scenarios where emotional training videos are scarce or nonexistent. Project website: https://yuangan.github.io/eat/

1. Introduction

Recently, there has been increasing attention on synthesizing realistic talking heads due to their wide-ranging applications in industry, such as digital human animation [20, 22, 55], visual dubbing [43], and video content creation [50]. Audio-driven talking-head generation aims to produce realistic talking-head videos synchronized with speech. However, unlike speech, humans convey intentions through emotional expressions. Therefore, generating emotional talking heads is important to improve the fidelity of talking heads for real-world applications. To address this open problem, various forms of knowledge (such as human head models, emotions, audio, and vision) must be considered in constructing multi-knowledge representations [57].

Previous one-shot talking-head generation methods [67, 43, 55] focus on achieving audio-visual synchronization for emotion-agnostic talking-heads, which is a special case of realistic talking-heads. More recent works [23, 22, 34] pay attention to generating emotion-aware talking-heads. GC-AVT [34] and EAMM [22] are two methods that generate emotional videos using driven emotional videos and pose-guiding videos. GC-AVT [34] achieves explicit control over the expression, speech content, and pose of talking heads through granular pre-processing design. EAMM [22] synthesizes one-shot emotional talking-heads by adding augmented emotional source videos. Since the driven videos can introduce semantic ambiguity to mouth shapes, GC-AVT replaces the mouth part with the neighbor frame, while EAMM ignores the mouth part of driven emotional videos by using data augmentation. Additionally, these methods require training or finetuning the entire network at high costs for emotional talking-heads generation.

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{fig1.png}
\caption{Efficient emotional talking-head generation. (a) Previous work trains or finetunes the whole network with augmented emotional driving videos. (b) Our EAT transforms emotion-agnostic talking-head models into emotion-controllable ones through flexible guidance, including emotional prompts or text-guided CLIP [45] supervision, by lightweight adaptations.}
\end{figure}
Although emotion-aware methods have made progress in the one-shot talking-head generation, they lack in-depth thinking in two key aspects. (1) **Architecture efficiency.** As a sub-task of talking-head generation, it is parameter-inefficient to train or finetune the entire emotional talking-head generation network. Furthermore, since large-scale emotion-agnostic talking-head data is more readily available than emotional data, it is worthwhile to consider how to efficiently reuse the knowledge learned from emotion-agnostic data. (2) **Guidance flexibility.** Previous methods prefer to transfer the driving videos to the target talking heads rather than directly learning the emotional representation. In practice, finding appropriate driven emotional videos requires taking into account factors such as resolution, occlusion, and even the length of the driven emotional videos and audio. Furthermore, prior research has neglected to consider lip shapes, which can result in unrealistic emotional expressions. For instance, according to FACS [18, 27], depressed lip corners are one of the key components of the sad expression.

To address the above limitations, a desirable approach should enable an efficient and flexible transfer of pretrained talking-head models to emotional talking-head generation tasks with lightweight emotional guidance, as illustrated in Fig. 1. There are two key advantages. Firstly, with the reused knowledge, we can readily and effortlessly apply the talking-head model to emotional talking-head generation tasks. Secondly, obtaining lightweight guidance is simpler and more adaptable in practical scenarios, such as text-guided zero-shot expression editing.

To realize the aforementioned paradigm, we propose an efficient Emotional Adaptation framework for audio-driven Talking-head (EAT) generation, which involves two stages. In the first stage, we enhance the unsupervised 3D latent keypoints representation [56] to capture emotional expressions. Then, we introduce the Audio-to-Expression Transformer (A2ET), which learns to map audio to enhanced 3D latent keypoints using large-scale talking-head datasets. In the second stage, we propose learnable guidance and adaptation modules for steering emotional expression generation. These include Deep Emotional Prompts for parameter-efficient emotional adaptation, a lightweight Emotional Deformation Network (EDN) for learning the emotional deformation of facial latent representation, and a plug-and-play Emotional Adaptation Module (EAM) for enhancing visual quality. Our approach enables rapid transfer of traditional talking-head models to emotional generation tasks with high-quality results and supports zero-shot expression editing with image-text models [45].

We conduct extensive experiments to assess the effectiveness of EAT on emotional talking-head generation. Compared to baseline competitors, EAT achieves superior performance without guiding emotional videos. Moreover, based on the pretrained talking-head model, we can attain state-of-the-art (SOTA) performance in 2 hours with only 25% training data. The results indicate that our method is capable of generating more realistic talking-head videos. And with only text descriptions of emotions, we can achieve zero-shot talking-head editing.

In summary, the main contributions of our work are listed below:

- Our study introduces a new two-stage paradigm, called EAT, for addressing emotional talking-head tasks. Our experiments demonstrate that this paradigm outperforms previous methods with respect to both emotion manipulation and video quality in one-shot talking-head generation tasks.
- Our proposed architecture includes deep emotional prompts, an emotional deformation network, and an emotional adaptation module. This design enables the efficient transfer from generating talking heads without emotional expression to generating talking heads with emotional expression.
- To the best of our knowledge, our study is the first to introduce flexible guidance for talking-head adaptation. By utilizing image-text models, we can achieve zero-shot expression editing of talking-head videos, surpassing the capabilities of previous methods.

### 2. Related Work

Previous works have achieved great performance in audio-visual synchronization. However, there remain some challenges such as efficient knowledge transfer based on a large-scale pretrained talking-head model. **Audio-driven Talking Head Generation.** Audio-driven talking-head generation [10, 51, 63, 7, 67, 20, 55, 64] based on deep learning has attracted lots of attention in recent years. Chen *et al.* [7] design a two-stage structure that leverages facial landmarks as an intermediate representation. MakeItTalk [67] generates one-shot talking heads based on disentangled speech and speaker. PC-AVS [64] generates arbitrary talking heads with pose control.

Following the groundbreaking work of transformers [52] in NLP [15, 46], recent works have achieved remarkable progress in video-related tasks, including action recognition [2, 19, 36], video segmentation [58, 59, 68, 41], cross-model understanding [32, 8], etc. In the talking-head field, AVCT [55] designs an audio-visual correlation transformer for generating talking-head videos. In this work, we enhanced 3D latent keypoints and applied transformers to generate more realistic talking heads. **Emotion-aware Talking Head Generation.** Emotional talking-head generation has been studied recently for realistic talking-head generation. Pumarola *et al.* [44] introduce
Figure 2. **Overview of EAT model.** (a) In the first stage, the Audio-to-Expression Transformer (A2ET) transfers latent source image representation, source audio and head pose sequences to 3D expression deformation. (b) In the second stage, the emotional guidance is injected into A2ET, Emotional Deformation Network (EDN) and Emotional Adaptation Module (EAM) for emotional talking-head generation, presented in dashed lines. (c) The RePos-Net takes the 3D source keypoints \( K_s \) and driven keypoints \( K_i \) to generate frames.

an unsupervised framework to generate facial videos with a specific expression. EVP [23] proposes emotional video portraits to produce more vivid results based on source video. However, the one-shot emotional talking-head generation has emerged recently. Sinha [49] generates pose-fixed emotional talking-heads with graph convolution. GC-AVT [34] train an emotion and pose controllable model with a granular pre-processing design. EAMM [22] synthesizes one-shot emotional talking heads with augmented emotional source videos. In our work, we achieve emotion control with efficient adaptation based on a pretrained talking-head model.

**Efficient Finetuning.** Efficient finetuning has been studied for knowledge transfer and many techniques have been proposed including residual adapter [47], bias tuning [5] and side tuning [61]. Recently, prompt [35] has attracted more attention in vision task. Inspired by prompt tuning in the language model, prompt tuning recently has been proposed in various visual tasks [45, 24, 66, 25, 65, 17, 33] for effectiveness and efficiency. CoOp [66] and VPT [25] utilize learnable prompt vectors and achieve better performance. Moreover, to increase generalizability, CoCoOp [65] designed a lightweight network to learn prompt vectors for each image. In our work, we introduce deep emotional prompts, an emotional deformation network and emotional adaptation modules to achieve efficient and effective emotion-related knowledge transfer.

3. **Method**

To avoid the expensive end-to-end training and finetuning of previous methods [22, 34], we propose a two-stage paradigm, the efficient Emotional Adaptation for audio-driven Talking-head (EAT) generation method. Firstly, we introduce the enhanced 3D latent representations and the emotion-agnostic talking-head pretraining using the Audio-to-Expression Transformer (A2ET). (Sec. 3.1) Secondly, we present a parameter-efficient emotional adaptation approach to quickly adapt the pretrained talking-head model for emotional talking-head tasks. This approach includes deep emotional prompts, an Emotional Deformation Network (EDN), and an Emotional Adaptation Module (EAM). (Sec. 3.2) At last, we introduce our training objectives in detail. (Sec. 3.3)

3.1. **Emotion-agnostic Pretraining**

Compared to inefficient end-to-end training for emotional talking-head [34, 22], we explore leveraging the
Therefore, we enhance the latent representation for more realistic expression elements, such as eyebrows, lip corners, etc. To serve that this kind of transfer fails to account for other facial features, such as mouth movements, we encode acoustic features \( A_{1:n} \) and acoustic features \( S_{1:n} \) as inputs. The audio semantic features \( S_{1:n} \) are extracted via a speech recognition model [1] from the MFCC features. To derive acoustic features \( A_{1:n} \), we design an audio encoder to encode the mel spectrogram extracted with 80 mel bins and 1025 frequency bins.

As shown in Fig. 2(a), given frame \( i \), we extract its semantic context features from \( 2w + 1 \) audio frames. Initially, speech features \( S_{i−w:i+w} \) and head pose features \( P_{i−w:i+w} \) are converted into speech tokens. The 6DoF of frame \( i \) is encoded into a pose token \( p \). The A2ET encoder takes these tokens as input. Subsequently, to capture subtle mouth movements, we encode acoustic features \( A_{i−w:i+w} \) and the latent source image representation with an audio encoder [31] and keypoint detector \( D_b \). These representations are fused to obtain acoustic tokens that are used by the A2ET decoder to output the feature of \( 2w + 1 \) tokens.

Expression Deformation Prediction. The expression deformation \( E_i \), composed of \( k \) 3D offsets, can be predicted.

3.1.2 Audio-to-Expression Transformer.

Since the 3D latent keypoints are specific to the source identity and more complex than 2D latent keypoints [48, 55, 22], predicting the 3D keypoints sequence directly is a challenging task. We notice that the facial expressions are mainly represented by the expression deformation \( E_i \) in 3D latent keypoints. Thus, the purpose of A2ET is to learn the audio-visual synchronized expression deformation, which is composed of audio-visual feature extraction and expression deformation prediction.

Audio-visual Feature Extraction. Previous work [55] generates emotion-agnostic talking heads with transformer and phonemes. However, training transformer requires a large dataset, and phoneme extraction is challenging in noisy or accented speech. To address these limitations, we train our A2ET model on a large dataset Voxceleb2 [11] and extract speech features \( S_{1:n} \) and acoustic features \( A_{1:n} \) as inputs. The audio semantic features \( S_{1:n} \) are extracted via a speech recognition model [1] from the MFCC features. To derive acoustic features \( A_{1:n} \), we design an audio encoder to encode the mel spectrogram extracted with 80 mel bins and 1025 frequency bins.

These modifications enhance the representation ability of the learned 3D latent keypoints, which are the objective of our A2ET model.

3.1.1 Enhanced Latent Representation.

Given a talking-head frame \( i \), the 3D latent keypoints \( K_i \), which are learned unsupervisedly, are composed of four components: identity-specific canonical keypoints \( K_c \), rotation matrix \( R_i \), translation \( T_i \), and expression deformation \( E_i \). These components are then combined with the following equation:

\[
K_i = R_i K_c + T_i + E_i. \tag{1}
\]

Based on 3D latent keypoints, RePosition Network (RePos-Net) [56] can transfer facial expressions from one person to another, as illustrated in Fig. 3. However, we observe that this kind of transfer fails to account for other facial expression elements, such as eyebrows, lip corners, etc. Therefore, we enhance the latent representation for more realistic talking-head generation with these modifications:

- We remove the deformation prior loss in OSFV [56] which penalizes the magnitude of the keypoints deformation. This allows our latent keypoints to capture more subtle changes in facial expressions.
- We use the MEAD dataset [54] to acquire labeled and paired face data from neutral and emotional videos of the same identity. This helps the network learn more expressive faces from expression changes.
- To avoid the influence of expression-irrelevant background, we only compute losses on the facial part. And we augment our paired data with the affectnet [40] dataset to improve generalizability.
from the feature of central frame $i$, but optimizing the 3D expression motions leads to convergence issues. We observed that the 3D keypoints learned by self-supervision exhibit inherent inter-dependencies, with only a few keypoints influencing facial expressions. To address this, the principal component analysis (PCA) of $E_i$ is adopted to reduce the dimensionality and eliminate nonessential information. Then we can predict the 3D expression deformation from the audio features.

3.2. Efficient Emotional Adaptation

Traditional audio-driven talking-head methods [7, 67, 20, 55, 64] have made impressive advancements in emotion-agnostic talking-head generation tasks. However, to achieve realistic talking heads, emotional expression is crucial. Hence, we present a novel approach that includes three parameter-efficient modules for swift emotional adaptation from emotion-agnostic models. These modules comprise Deep Emotional Prompts, the Emotional Deformation Network (EDN), and the Emotional Adaptation Module (EAM), specifically designed to enable efficient emotional adaptation of pre-trained A2ET. Our approach allows for lightweight adaptations, which offer flexibility in guiding downstream tasks, such as zero-shot expression editing.

**Emotional Guidance.** One straightforward idea is to generate emotional talking heads using learnable guidance conditioned on emotions. We posit that each emotional type belongs to a distinct sub-domain in latent space. As shown in Fig. 2 (b), a mapping network $M$ is adopted to extract emotion-conditional guidance with a latent code $z \in U^{16}$. The latent code is sampled from a Gaussian distribution, which is commonly used in generative models [29, 9]. This emotional guidance is applied to steer the generation of emotional expressions.

**Deep Emotional Prompt.** To achieve parameter-efficient emotional adaptation, we include the guidance as an additional input token of the A2ET transformer layer, as shown in Fig. 2(a). We separately introduce shallow and deep emotional prompts into the A2ET transformer architecture, with the shallow prompt added to the first layer and the deep prompt added to every layer thereafter. Our results in Table 4 demonstrate that the deep prompt leads to better emotional expression transfer compared to the shallow prompt. However, we also observe that incorporating emotional prompts can have a detrimental effect on audio-visual synchronization. Generating emotional expressions using fixed transformer weights while ensuring audio-visual synchronization may pose a challenge for the prompts.

**Emotional Deformation Network.** We observe that the decoupled 3D implicit representations in Eq. 1 exhibit linear additivity. Furthermore, emotional talking heads show emotional deformation that is not present in traditional talking-heads. To complement $E_i$, one intuitive approach is to include an emotional expression deformation term:

$$E'_i = E_i + \Delta E_i,$$

where $E'_i$ represents emotional expression deformation, $E_i$ represents speech-related expression deformation predicted by A2ET, and $\Delta E_i$ represents emotion-related expression deformation. To predict $\Delta E_i$, we design a sub-network called the Emotional Deformation Network (EDN), which is depicted in Fig. 2(b). EDN utilizes the A2ET encoder architecture to predict $\Delta E_i$ with emotional guidance and the source latent representation token. To accelerate adaptation, we initialized EDN with the pretrained A2ET encoder. To update $E_i$ with $E'_i$, we can get emotional 3D latent keypoints using Eq. 1.

**Emotional Adaptation Module.** To enhance the visual
Once we have obtained $\gamma$, $\beta$ by $\gamma, \beta = \tanh(FC(ReLU(FC(e))))$. (3)

Once we have obtained $\gamma$ and $\beta$, we can input the feature $x$ to obtain the emotional feature, which is calculated using the following equation:

$$EAM(x) = F_s(1 + \gamma, x) + \beta,$$ (4)

where $F_s$ denotes the channel-wise multiplication. As shown in Figure 2, the EAM can be inserted into RePoseNet, as well as the audio and image feature extractors.

### Zero-shot Expression Editing

Owing to the quick adaptation capabilities of our EAT, we can achieve zero-shot text-guided expression editing of talking heads by distilling knowledge from CLIP [45], a large-scale vision-language pertaining model. This unique ability sets our work apart from the latest research [22], as it eliminates the need for emotional training data and enables generalization to applications that require rare expressions.

Specifically, our goal is to employ CLIP loss to learn emotional guidance correlated with text-described expressions. To achieve this, we extract the head poses, source audio and the first frame from the target video as the input. Besides, a target expression description is taken for finetuning. Utilizing the refined EAT model and our training loss, we add an additional CLIP loss [42] to finetune the mapping network and EAM module only. In detail, we extract image embeddings from the predicted talking face using the image encoder of CLIP, and text embeddings from the description using its text encoder. We then iteratively optimize the distance between the image and text embeddings to align the generated talking face with the input text.

### 3.3. Training Objectives

For supervised learning, the loss is calculated as follows:

$$L = \lambda_{lat} L_{lat} + \lambda_{sync} L_{sync} + \lambda_{rec} L_{rec},$$ (5)

where $\lambda_{lat}$, $\lambda_{sync}$ and $\lambda_{rec}$ are hyper-parameters that re-weight the corresponding term. As for zero-shot editing, we replace the $\lambda_{rec} L_{rec}$ with the CLIP loss since there is no ground-truth video. In the following, we will discuss each training loss in detail.

#### Latent Loss

Latent loss is applied to optimize the predicted latent keypoints:

$$L_{lat} = \frac{1}{N} \sum_{i=1}^{N} (\|PE_i - P\hat{E}_i\|^2_2 + \|K_i - \hat{K}_i\|^2_2),$$ (6)

where $N$ denotes the frame length of a sampled audio clip in each batch. $PE_i$ represents the predicted PCA of expression deformation in $i$ frame. $K_i$ is the transferred 3D latent keypoints according to Eq. 1. $P\hat{E}_i$ and $\hat{K}_i$ are the corresponding ground-truth of the frame $i$. As the emotional expression deformation $\Delta E_i$ is not included in $P_i$, we only use the loss of 3D keypoints in Eq. 6 while training EDN.
Sync Loss. The synchronization loss is introduced in Wav2Lip [43]. Based on the structure of SyncNet [13], we train an expert to discriminate the audio-visual synchronization in neutral and emotional datasets. For a sampled audio clip in each batch, we computed the synchronization loss of the generated video using the following equation:

$$L_{\text{sync}} = -\log\left(\frac{v \cdot s}{\max(\|v\|_2 \cdot \|s\|_2, \epsilon)}\right).$$  \hspace{1cm} (7)

The input speech embedding $s$ and the generated video embedding $v$ are extracted by the speech encoder and image encoder in SyncNet, respectively.

Reconstruction Loss. To improve the expression generation, we employ $L_1$ reconstruction loss only in the facial region. Additionally, to generate sharper frames, we apply perceptual loss [26] to the entire frames using a pretrained VGG19 model.

CLIP Loss. The CLIP loss calculates the similarity between the embeddings of the generated face and the text description with a pretrained CLIP model. Specifically, the CLIP loss is calculated by taking the cosine similarity between the normalized embedding of the image and text.

4. Experiments

4.1. Experimental Setup

Implementation details. The videos are sampled to 25 FPS and the audio sample rate is 16KHz. The videos are cropped and resized to $256 \times 256$. To synchronize the audio features and the video, we extract the mel-spectrogram [37] by configuring the window length and hop length to 640. The number of keypoints $k$ used in EAT is 15. The mapping network of emotional prompts consists of shared four MLP layers and unshared four MLP layers for every kind of emotion. We enhance the 3D latent keypoints for 48 hours and pretrain the A2ET with enhanced latent keypoints for 48 hours. Then we finetune the EAT architecture for 6 hours. Our work is based on 4 NVIDIA 3090 GPUs.

Datasets. The training dataset consists of videos from VoxCeleb2 [11] and MEAD [54]. MEAD is a high-quality emotional talking-head video set with 8 kinds of emotions. To ensure fair comparisons, we split the MEAD dataset into training and testing sets based on identity, using the same test identities as EAMM [22]. To learn large head pose changes, we select about 8,000 emotional videos from the VoxCeleb2 [11] dataset with Emotion-FAN [39] for finetuning. To obtain the PCA of the enhanced 3D keypoints, we extract the largest 32 eigenvalue matrix and mean of 2,500 videos from the training set.

Baselines. We compare with SOTA one-shot talking-head generation methods on LRW [12] and MEAD [54] test set. They are ATVG [7], Wav2Lip [43], MakeItTalk [67], AVCT [55], PC-AVS [64] and EAMM [22].

Metric. We assess the quality of synthesized emotional videos with the following metrics:

Image quality. We utilize PSNR, SSIM and Frechet Inception Distance score (FID) [21] to measure the image quality of synthesized videos.

Audio-visual synchronization. We evaluate the audio-visual synchronization of the synthesized videos using the confidence score of SyncNet [13]. In addition, the distance between the landmarks of the mouth (M-LMD) [7] is used to indicate speech content consistency, while the distance between the landmarks of the whole face (F-LMD) represents the accuracy of the pose and expressions.

Emotional accuracy. To assess the emotional accuracy (Acc emo) of the generated emotions, we fine-tune the Emotion-Fan [39] using the training set of MEAD.

4.2. Talking-head Generation

To verify the effectiveness of EAT, we do experiments on emotion-agnostic and emotional talking-head generation.

Emotion-agnostic Talking-head Generation. For one-shot emotion-agnostic talking-head generation, we test on the LRW test set, which comprises 25k neutral videos. We take the first frame as the source image for every test video. As presented in Table 1, our method outperforms other...
methods in terms of visual quality for emotion-agnostic talking-head generation. Furthermore, our EAT can improve the performance of the pretrained talking-head model. Wav2Lip and PC-AVS risk overfitting the pretrained lip synchronization scoring model since their sync scores surpass the ground truth. Additionally, Wav2Lip only generates mouth parts without facial expression and head pose.

**Emotional Talking-head Generation.** We follow the setting of EAMM to compare the emotional talking-head generation on the public-available MEAD test set. For all methods in terms of visual quality for emotion-agnostic talking-head generation. Furthermore, our EAT can improve the performance of the pretrained talking-head model. Wav2Lip and PC-AVS risk overfitting the pretrained lip synchronization scoring model since their sync scores surpass the ground truth. Additionally, Wav2Lip only generates mouth parts without facial expression and head pose.

To validate the effectiveness of our enhanced latent representation, we generate emotional videos driven by source images and emotional videos in the MEAD test set.

Table 3. Ablation study of enhanced latent representation. To validate the effectiveness of our enhanced latent representation, we generate emotional videos driven by source images and emotional videos in the MEAD test set.

<table>
<thead>
<tr>
<th>Prompt</th>
<th>PSNR↑</th>
<th>M/F-LMD↓</th>
<th>Sync↑</th>
<th>Acc emo↑</th>
</tr>
</thead>
<tbody>
<tr>
<td>w/o</td>
<td>20.46</td>
<td>2.85/2.99</td>
<td>8.12</td>
<td>25</td>
</tr>
<tr>
<td>Shallow</td>
<td>21.19</td>
<td>2.50/2.63</td>
<td>7.63</td>
<td>57</td>
</tr>
<tr>
<td>Deep</td>
<td>21.23</td>
<td>2.36/2.48</td>
<td>7.83</td>
<td>84</td>
</tr>
</tbody>
</table>

Table 4. Ablation study of prompt. To verify the effectiveness of shallow and deep emotional prompts, we produce videos using source images and emotional videos from the MEAD test set.

4.3. User Study

To evaluate the generated emotional talking heads with in-the-wild images, we conducted a user study with 14 participants to assess lip-sync, video quality, and emotion classification. The maximum score value is 5. To ensure a diverse set of images, we randomly sampled 16 images from both the CREMA-D and LRW datasets, which are not included in the training data. Additionally, we used audio from MEAD test set to generate a total of 32 videos (4 x 8 emotions) for each method. As shown in Table 2, our method achieved the best scores for lip-sync, video quality, and emotion accuracy, validating the effectiveness of our EAT architecture in generating emotional talking heads. Besides, the results show that the sync value in Table 1 is inaccurate for emotional talking-head due to the SyncNet model being trained only with neutral talking-head videos. More in-the-wild results can be found in the supplementary.

4.4. Tuning Efficiency

During the finetuning of the second stage, we conducted periodic tests on a subset of the MEAD test set every half hour. This enabled us to demonstrate the time and data efficiency of our EAT. EAT can efficiently adapt the pretrained
Table 5. Ablation study of each component. Each component contributes to the improvement of the video quality, thus verifying its effectiveness.

<table>
<thead>
<tr>
<th>Ablation</th>
<th>(A)</th>
<th>(B)</th>
<th>(C)</th>
<th>EAT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prompt</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>EDN</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>EAM</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>PSNR↑</td>
<td>20.46</td>
<td>21.23</td>
<td>21.40</td>
<td>21.79</td>
</tr>
<tr>
<td>M/F-LMD↓</td>
<td>2.85/2.99</td>
<td>2.36/2.48</td>
<td>2.28/2.41</td>
<td>2.22/2.43</td>
</tr>
<tr>
<td>Sync↑</td>
<td>8.12</td>
<td>7.83</td>
<td>7.83</td>
<td>8.22</td>
</tr>
<tr>
<td>Acc emo↑</td>
<td>25</td>
<td>84</td>
<td>81</td>
<td>67</td>
</tr>
</tbody>
</table>

Table 6. Ablation study of each loss. Each loss contributes to the improvement of the video quality or synchronization value.

<table>
<thead>
<tr>
<th></th>
<th>(A)</th>
<th>(B)</th>
<th>(C)</th>
<th>EAT</th>
</tr>
</thead>
<tbody>
<tr>
<td>L_per</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>L_lat</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>L_sync</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>L_1</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>PSNR↑</td>
<td>21.52</td>
<td>21.61</td>
<td>21.31</td>
<td>21.79</td>
</tr>
<tr>
<td>Sync↑</td>
<td>5.50</td>
<td>5.66</td>
<td>8.13</td>
<td>8.22</td>
</tr>
</tbody>
</table>

Enhanced Latent Representation. We compare the latent representation before and after our proposed enhancement. As shown in Table 3, the superior performance in face reenactment shows that our pre-trained model can capture a wider range of emotional facial movements than OSFV [56], which has been discussed in Sec. 3.1.1.

Prompt. To study the effect of different types of prompts, we conduct experiments with shallow and deep emotional prompts. Table 4 shows that deep prompt learns emotional expression deformation better than shallow prompt, although it has a side effect on synchronization.

Each Component. To verify the effectiveness of our proposed modules, we conduct ablation experiments by removing one component every time. Table 5 shows that all three components can improve video quality. Deep emotional prompts can transfer the talking-head knowledge to emotional talking-head generation at the expense of synchronization. Although deep emotional prompts lead to intense emotional expressions, the outputs deviate from the ground truth. By incorporating EDM and EAM, the image quality and fidelity toward the ground truth are enhanced, as evidenced by the rise in PSNR/SSIM values. Yet, this comes at the expense of emotion intensity and accuracy. Please refer to our supplementary for visual analysis.

Each Loss. As shown in Table 6, we conduct an ablation study on perceptual loss, latent loss, sync loss, and L_1 loss in the finetuning stage. It demonstrates that sync loss contributes to synchronization while others contribute to expression fidelity.

5. Conclusion

In this paper, we propose an efficient emotional adaptation paradigm for audio-driven talking-head generation, consisting of two stages. First, we enhance 3D latent representation and develop a transformer architecture A2ET to achieve emotion-agnostic talking-head generation. Second, we introduce learnable guidance for emotional expression control through our deep emotional adaptation modules, EDM, and EAM. With these adaptation modules, EAT can quickly transfer the pretrained talking-head model to emotional talking-head generation. Experiments demonstrate that our EAT is the first parameter-efficient and effective paradigm for emotional talking-head generation.

Limitations and Broader Impact. 1) The drawbacks of emotional training data, such as the diversity of background and head poses, will affect the generalizability of our EAT. 2) Our approach paves the way for broader talking-head applications, including zero-shot or one-shot emotional talking-head generation.

Acknowledgements. This work is supported by the National Key R&D Program of China (2022YFB3303300) and the Fundamental Research Funds for the Central Universities (No. 226-2023-00048).
References


Supplementary Material

A. The Networks Details

We provide additional details of our network architecture and training procedure. It should be noted that the Keypoint Detector ($D_D$) and RePos-Net networks are primarily derived from OSFV [56]. For more information, interested readers may refer to OSFV [56].

**Audio-to-Expression Transformer.** We use the Audio-to-Expression Transformer (A2ET) to transfer the audio to 3D latent expression deformation sequences. The A2ET consists of an encoder and a decoder, both with 6 transformer layers and 8 heads. The feed-forward layer has a dimension of 1024. Each token is a 128-dim vector. The expression deformation vector ($E_D$) is predicted by the feature of the central frame $i$. However, directly optimizing the 3D expression motions leads to convergence problems in network training. To address this issue and bridge the gap between the 3D expression deformation and the audio features, we use principal component analysis (PCA) to reduce the dimensionality of $E_D$ from 45 to 32. Specifically, we calculate $E_D = \text{proj}(f_D, U, M)$, where $P$ serves as the emotional guidance for the emotional adaptation module (EAM), which modifies the features of the audio-to-expression transformer (A2ET). The remaining six tokens $e_1, \ldots, e_6$ are fed to the corresponding transformer layer of A2ET as deep emotional prompts. The Emotional Deformation Network (EDN) then uses all these tokens and the latent source representation to generate the emotional deformation $\Delta E$.

**Emotional Deformation Network.** The Emotional Deformation Network (EDN) learns the emotional deformation $\Delta E$ using the same architecture as the A2ET encoder, which has six transformer layers. Fig. 10(b) shows the input and output of EDN. It takes the latent source representation $d$ and the emotional guidance tokens $e_0, e_1, \ldots, e_6$ as input, and extracts their features $f_{d0}, f_{e0}, \ldots, f_{e6}$. Then it applies global average pooling to the emotion-related features $f_{d0}, f_{e0}, \ldots, f_{e6}$ and uses an MLP layer to obtain the final emotional deformation $\Delta E$.

**Emotional Adaptation Module.** The emotional adaptation module (EAM) consists of two multi-layer perceptrons (MLPs). As shown in Fig. 10(c), given the input feature $x$ and the emotional token $e_0$, we extract the weight vector $f_0$, $f_1$, $\ldots$, $f_6$ as input, and uses an MLP layer to obtain the output $x'$.

**Parameter Efficiency Analysis.** Our Deep Emotional Prompts, EDN and EAM only require about 7% of the parameters compared to the whole network. The emotion mapper, which generates deep emotional prompts for eight emotions, has most of the parameters. In addition, EDN and EAM consume less than 2%. These parameters are 13.9M. This is half of the emotional network of EAMM [22], which has 27.9M parameters.

B. Training and Testing Details

**Training Details.** We use the MEAD dataset and 8k emotional video clips from Voxceleb2 [11] with various facial expressions to learn the enhanced latent keypoints. We also use roughly 21k emotional images from AffectNet [40] to train emotional expression generation. Due to the lack of corresponding neutral faces, we generate neutral faces paired with emotional images by using Ganimation [44]. We train our EAT with Adam [28] with $\beta_1 = 0.5$ and $\beta_2 = 0.999$. The learning rate is set to $1.5 \times 10^{-4}$ for A2ET and $2 \times 10^{-4}$ for other modules. In the first stage, we train A2ET with only latent loss first to obtain a good initialization, and then we train it with full loss. To improve gener-
alization, we use the Voxceleb2 and MEAD datasets, which contain about 225k video clips. In the second stage, we finetune efficient adaptation modules with only the MEAD dataset, which has about 10k video clips. We test our model on LRW [12] and MEAD [54] dataset.

**Testing Details and Protocol.** When testing LRW, the input is the first frame, and the transformation starts from the first frame. Therefore, the relative offsets of the latent keypoints are used. When testing MEAD, due to the variation in facial expressions, which is unrelated to the neutral source image, the predicted latent keypoints are used.

To ensure accurate evaluations, we crop and align [7] the faces before calculating these metrics: PSNR, SSIM, FID, M-LMD, and F-LMD. As for synchronization confidence, we preprocess the generated videos with reference to PC-AVS [64].

### C. Additional Experimental Results

**Additional baseline results.** As shown in Figure 11, we compare our EAT results with several baseline methods. Our results are more pleasant than those of MakeItTalk [67] and Wav2Lip [43], which do not consider emotional expression in talking heads. Additionally, our EAT achieves emotion control compared to the pretrained A2ET network. Videos are included in the supplementary material for reference.

**Various emotional expressions** To validate the diversity of emotional expressions generated by EAT, we present six different emotional results in Fig. 9. Compared to Neutral emotion, emotional expressions result in different modifications to facial elements, such as lip corners, eyes, and brows. We present the quantitative results of emotion classification in Table 7. We notice that EAT works significantly better on Happy, Sad, Disgusted, and Contempt than other methods. This is because our method can capture mouth details and these emotions can be more clearly reflected by the lips. As for Neutral, Angry, and Surprised, EAMM [22] performs well since these emotions are more apparent on the eyes and brows. And EAT can also achieve better performance in these emotions. However, all methods perform poorly on Fear emotion. It may be because Fear and Surprise are similar and diffi-

---

<table>
<thead>
<tr>
<th>Weight</th>
<th>PSNR↑</th>
<th>M/F-LMD↓</th>
<th>Sync↑</th>
<th>Acc\textsubscript{emo}↑</th>
</tr>
</thead>
<tbody>
<tr>
<td>w/o</td>
<td>21.49</td>
<td>2.27/2.46</td>
<td>8.02</td>
<td>76</td>
</tr>
<tr>
<td>EAT</td>
<td>21.79</td>
<td>2.22/2.43</td>
<td>8.22</td>
<td>67</td>
</tr>
</tbody>
</table>

Table 8. Ablation study of EDN weight initialization. The weight initialization of EDN with the A2ET encoder promotes the performance of EAT.

---

<table>
<thead>
<tr>
<th>Method</th>
<th>Happy</th>
<th>Angry</th>
<th>Disgusted</th>
<th>Fear</th>
<th>Sad</th>
<th>Neutral</th>
<th>Surprised</th>
<th>Contempt</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wav2Lip [43]</td>
<td>0.00</td>
<td>25.64</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>91.25</td>
<td>0.00</td>
<td>0.00</td>
<td>17.87</td>
</tr>
<tr>
<td>MakeItTalk [67]</td>
<td>0.00</td>
<td>25.64</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>75.00</td>
<td>0.00</td>
<td>0.00</td>
<td>15.23</td>
</tr>
<tr>
<td>AVCT [55]</td>
<td>0.83</td>
<td>25.64</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>69.38</td>
<td>0.00</td>
<td>10.08</td>
<td>15.64</td>
</tr>
<tr>
<td>EAMM [22]</td>
<td>23.33</td>
<td>84.48</td>
<td>9.40</td>
<td>0.00</td>
<td>0.00</td>
<td>98.13</td>
<td>94.02</td>
<td>72.27</td>
<td>49.85</td>
</tr>
<tr>
<td>Pretrain (Ours)</td>
<td>35.00</td>
<td>11.97</td>
<td>0.00</td>
<td>0.00</td>
<td>49.17</td>
<td>38.75</td>
<td>0.00</td>
<td>59.66</td>
<td>25.18</td>
</tr>
<tr>
<td>EAT (Ours)</td>
<td>84.17</td>
<td>100.00</td>
<td>48.72</td>
<td>16.52</td>
<td>49.17</td>
<td>100.00</td>
<td>100.00</td>
<td>94.96</td>
<td>75.43</td>
</tr>
</tbody>
</table>

Table 7. Quantitative evaluation of the emotion classification in the MEAD dataset.
Figure 11. More Qualitative results. We compare with more baselines, such as MakeItTalk [67], Wav2Lip [43], and our pretrained model.

Figure 12. Emotion interpolation. The top row is the emotion interpolation results between Angry and Surprised. The bottom row is the results between Contempt and Sad. The neutral faces are from MEAD [54].

Emotion interpolation. We conduct emotional guidance interpolation on the MEAD test set to verify that the latent space learned by the emotion mapper is continuous, as Fig. 12 shows.

Additional ablation study. We conduct further ablation studies on the weight initialization of EDN. Our results, presented in Table 8, show that using the weight initialization of the A2ET encoder leads to quicker convergence and improved performance in terms of video quality and audio-visual synchronization.

Visual analysis on each component of EAT. To analyze the effect of each component of our model, we show the fear emotion results from (A), (B), and (C), with corresponding accuracy rates of 38.46%, 30.77%, and 15.38% respectively in Fig. 13. Deep emotional prompts help generate intense emotional expressions that deviate from the Ground Truth (GT). By incorporating EDM and EAM, we achieve greater fidelity toward the GT and higher image quality in terms of PSNR/SSIM. This is attributed to the learning capabilities of EDM and EAM for emotional data. However, it results in reduced emotion intensity and accuracy.

Visualization on the profile faces. To assess the ability of enhanced latent representation in 3D talking-head generation, as shown in Fig. 14, we visualize the talking-head frames generated from the profile faces of MEAD. We test the faces captured from left 30 degrees and right 60 degrees with Surprised and Happy emotions.

D. Limitations and Future Work.

While EAT is capable of generating emotional talking-head videos with emotional guidance, there are still some
### MEAD profile Results

<table>
<thead>
<tr>
<th>Source</th>
<th>(A) Prompt</th>
<th>(B) Prompt+EDM</th>
<th>(C) Prompt+EDM+EAM</th>
<th>GT</th>
</tr>
</thead>
<tbody>
<tr>
<td>PSNR/SSIM</td>
<td>22.37/0.7022</td>
<td>23.22/0.7399</td>
<td>23.99/0.7460</td>
<td>∞/1.0000</td>
</tr>
</tbody>
</table>

#### Figure 13. Visualization on each component of EAT.

#### Figure 14. Visualization on the profile faces of MEAD.

#### Figure 15. Failure case. The driving audio and poses are from the videos in the first row. The second and third rows display the generated results with Surprised emotional guidance. Neutral faces are from MakeItTalk [67] and driving video is from LRW [12].

**E. Ethical Considerations.**

Our research is intended for use in virtual human research and entertainment. However, there is a risk that the emotional talking-head generation algorithm could be abused. We strongly recommend that generated talking-head videos be labeled as “fake”. On one hand, our work demonstrates that emotional talking-head generation is technically feasible. On the other hand, fake video detection [14, 38] has attracted significant attention. We would be happy to assist in the development of related research.

Limitations. Firstly, the diversity of background and head pose in emotional training data can affect the generalizability of our EAT. As shown in Fig. 15, the wrinkles on the forehead are not obvious in these in-the-wild images. This issue could be addressed by more naturalistic and non-acted emotional data [60, 30, 4] and representations with the head prior, such as FLAME[62]. Secondly, effective guidance texts are required to achieve zero-shot generation. This may be due to the limited ability of models trained on image-text pairs to capture emotional expression, which could affect the performance of zero-shot learning. Thirdly, the eye region, such as eye blinks [53] and gaze direction [16], has not been considered in our work. Finally, the discrete emotion guidance limits the representation ability of our model. It needs to note that facial expressions are not always representative of the internal emotional state [3]. More refined theories of emotion, such as the valence-arousal model, may help generate more realistic emotions. We leave these problems for future work.