EDTalk: Efficient Disentanglement for Emotional Talking Head Synthesis (Supplementary Material)

Shuai Tan[1], Bin Ji[1], Mengxiao Bi[2], and Ye Pan[1]*

[1] Shanghai Jiao Tong University
[2] NetEase Fuxi AI Lab
{tanshuai0219, bin.ji, whitneypanye}@sjtu.edu.cn
bimengxiao@corp.netease.com

In the main paper, we introduce an innovative framework designed to produce emotional talking face videos, which enables individual manipulation of mouth shape, head pose, and emotional expression, conditioned on both video and audio inputs. This appendix delves deeper into: 1) Implementation Details. 2) Additional Experimental Results. 3) Discussion. In addition, we highly encourage viewing the Supplementary Video.

1 Implementation Details

1.1 Network Architecture

We utilize identical structures for Generator G in LIA [28]. We recommend consulting their original paper for further elaboration. Here, we delineate the details of the other network architectures depicted in Fig. 1.

Encoder E. The component projects the identity source I^i and driving source I^* into the identity feature f^{id} and the latent features $f^{i\to r}$, $f^{*\to r}$. It comprises several convolutional neural networks (CNN) and ResBlocks. The outputs of ResBlock serve as the identity feature f^{id}, which is then fed into Generator G to enrich identity information through skip connections. Subsequently, four multi-layer perceptrons (MLP) are employed to generate the latent features $f^{i\to r}$, $f^{*\to r}$.

MLP^m, MLP^p, MLP^e and MLP^m_A. To achieve efficient training and inference, these four modules are implemented with four simple MLPs.

Audio Encoder E_a. This network takes audio feature sequences $a_{1:T}$ as input. These sequences are passed through a series of convolutional layers to produce audio feature $f^a_{1:N}$.

* Corresponding author.

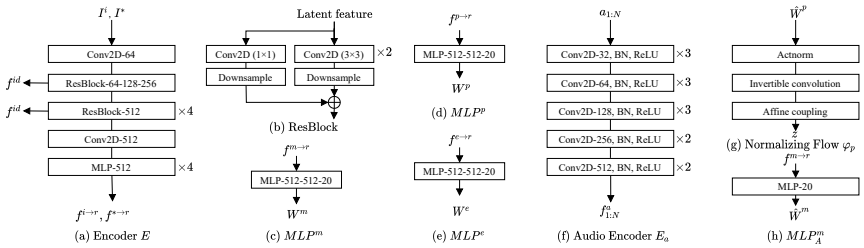

Fig. 1: Detailed architecture for different components in our EDTalk.

Normalizing Flow φ_p. Normalizing Flow φ_p comprises K flow step, each consisting of actnorm, invertible convolution and the affine coupling layer. Initially, given the mean μ and standard deviation δ for the weights W^p of pose bank B^p, actnorm is implemented as an affine transformation $h' = \frac{\beta-\mu}{\delta}$. Subsequently, φ_p introduces an invertible 1×1 convolution layer, $h'' = \mathbf{W} \cdot h'$, to handle potential channel variable. Following this, we utilize a transformer-based coupling layer $\mathcal{F}$ to derive z from h'' and $f^a_{1:N}$. Specifically, we split h'' into h''_{h1} and h''_{h2}, where h''_{h2} undergoes affine transformation by $\mathcal{F}$ based on h''_{h1}: $t, s = \mathcal{F}(h''_{h1}, f^a_{1:N}); h = (h''_{h2} + t) \odot s$, where t and s represent the transformation parameters. Thanks to the unchanged h''_{h1}, tractability is easily maintained in reverse. In summary, we can map W^p into the latent code z and predict weight $\hat{W}^p$ from a sampled code $\hat{z} \in p_{\mathcal{Z}}$ as follows:

$$z = \varphi_p^{-1}(W^p, f^a_{1:N}) \tag{1}$$

$$\hat{W}^p = \varphi_p(\hat{z}, f^a_{1:N}) \tag{2}$$

1.2 Data Details

Datasets

MEAD. MEAD entails 60 speakers, with 43 speakers accessible, delivering 30 sentences expressing eight emotions at three varying intensity levels in a laboratory setting. Consistent with prior studies [5, 9], we designate videos featuring speakers identified as 'M003,' 'M030,' 'W009,' and 'W015' for testing, while the videos of the remaining speakers are allocated for training.

HDTF. The videos of the HDTF dataset are collected from YouTube, renowned for their high quality, high definition content, featuring over 300 distinct identities. To facilitate training and testing, we partition the dataset using an 8:2 ratio based on speaker identities, allocating 80% for training and 20% for testing.

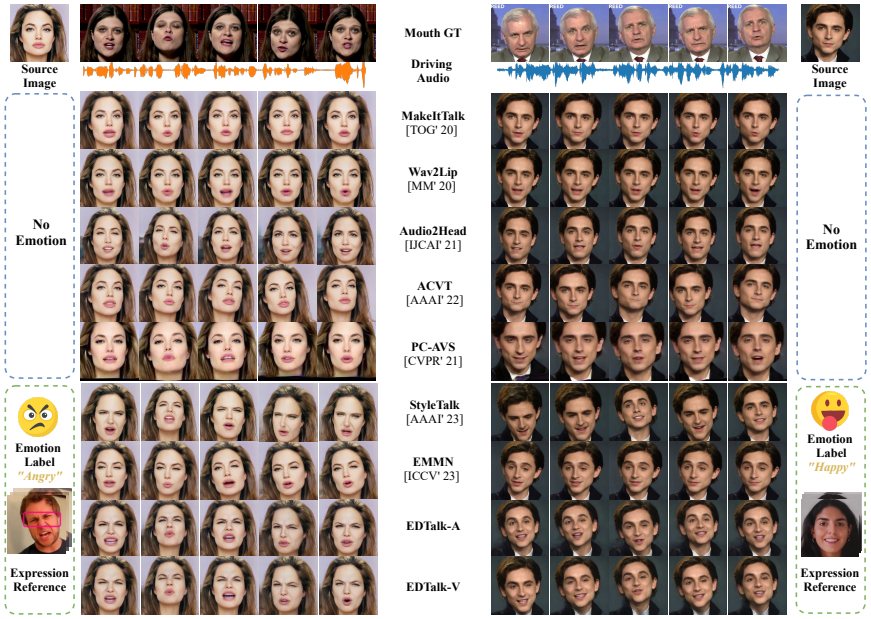

Fig. 2: Additional qualitative results, which are supplement to the main paper

Voxceleb2. Voxceleb2 [3] is a large-scale talking head dataset, boasting over 1 million utterances from 6,112 celebrities. It's important to note that we solely utilize Voxceleb2 for evaluation purposes, selecting 200 videos randomly from its extensive collection.

LRW. LRW [4] is a word-level dataset comprising more than 1000 utterances encompassing 500 distinct words. For evaluation, we randomly select 500 videos from the dataset.

Data Processing For video preprocessing, we employ face cropping and resize the cropped videos to the resolution of 256×256 for training and testing following FOMM [18]. Adhere to Wav2Lip [14], audio is down-sampled to 16 kHz and transformed into mel-spectrograms using an FFT window size of 800, hop length of 200, and 80 Mel filter banks. During the evaluation, for datasets without emotional labels, we utilize the first frame of each video as the source image and the corresponding audio as the driving audio to generate talking head videos. For emotional videos sourced from MEAD, we use the video itself as an expression reference. We select a frame with a 'Neutral' emotion from the same speaker as the source image for emotional talking head synthesis.

Method	Voxceleb2 [3]					LRW [4]				
	PSNR↑	SSIM↑	M-LMD↓	F-LMD↓	$\text{Sync}_{\text{conf}}$ ↑	PSNR↑	SSIM↑	M-LMD↓	F-LMD↓	$\text{Sync}_{\text{conf}}$ ↑
MakeItTalk [33]	20.526	0.706	2.435	2.380	3.896	22.334	0.729	2.099	1.960	3.137
Wav2Lip [14]	20.760	0.723	2.143	2.182	**8.680**	23.299	0.764	1.699	1.703	**7.545**
Audio2Head [25]	17.344	0.577	3.651	3.712	5.541	18.703	0.601	2.866	3.435	5.428
PC-AVS [32]	21.643	0.720	2.088	1.830	7.928	16.744	0.509	5.603	4.691	3.622
AVCT [26]	18.751	0.645	2.739	3.062	4.238	21.188	0.689	2.290	2.395	3.927
SadTalker [30]	20.278	0.700	2.252	2.388	6.356	-	-	-	-	-
IP-LAP [30]	20.955	0.724	2.125	2.154	3.295	**23.727**	0.770	1.779	1.683	3.027
TalkLip [22]	20.633	0.723	2.084	2.191	6.520	22.706	0.751	1.803	1.770	6.021
EAMM [9]	17.038	0.562	4.172	4.163	3.815	18.643	0.607	3.593	3.773	3.414
StyleTalk [12]	21.112	0.722	2.113	2.136	2.120	21.283	0.705	2.394	2.142	2.430
PD-FGC [21]	22.110	0.729	**1.743**	1.630	6.686	22.481	0.711	**1.576**	1.534	6.119
EAT [5]	20.370	0.689	2.586	2.383	6.864	21.384	0.704	2.128	1.927	6.630
EDTalk-A	**22.107**	**0.763**	1.851	**1.608**	6.591	23.409	**0.779**	1.729	**1.379**	6.914
EDTalk-V	**22.133**	**0.764**	1.829	**1.583**	6.155	**24.574**	**0.823**	**1.202**	**1.139**	6.027
GT	1.000	1.000	0.000/0.000	0.000	6.808	1.000	1.000	0.000/0.000	0.000	6.952

Table 1: Quantitative comparisons with state-of-the-art methods. We test each method on Voxceleb2 and LRW datasets, and the best scores in each metric are highlighted in bold. The symbol " ↑ " and " ↓ " indicate higher and lower metric values for better results, respectively.

1.3 Training Details

The encoder E and generator G are pre-trained in a similar setting as LIA [28]. Subsequently, we freeze the weights of the encoder E and generator G, focusing solely on training the Mouth-Pose Decouple Module. In this stage, our model is trained exclusively on the emotion-agnostic HDTF dataset, where videos consistently exhibit a 'Neutral' emotion alongside diverse head poses. It ensures that the Mouth-Pose Decouple Module concentrates solely on variations in head pose and mouth shape, avoiding the encoding of expression-related information. All loss function weights are set to 1. The training process typically requires approximately one hour, employing a batch size of 4 and a learning rate of 2e-3, executed on 2 NVIDIA GeForce GTX 3090 GPUs with 24GB memory. Once the Mouth-Pose Decouple Module is trained, we freeze all trained parameters and solely update the expression-related modules, including MLP^e, expression bases B^e, and the Expression Enhance Module EEM, utilizing both the MEAD and HDTF datasets. This stage typically takes around 6 hours, employing a batch size of 10 and a learning rate of 2e-3, conducted on 2 NVIDIA GeForce GTX 3090 GPUs with 24GB memory. We train our Audio-to-Lip model on the HDTF dataset for 30k iterations with a batch size of 4, requiring approximately 7 hours of computation on 2 NVIDIA GeForce GTX 3090 GPUs with 24GB memory. The Audio-to-Pose model is trained on the HDTF dataset for one hour.

2 Additional Experimental Results

2.1 More Comparison with SOTA Audio-Driven Talking Face Generation Methods

More quantitative results. Apart from the quantitative assessments conducted on the MEAD and HDTF datasets, as detailed in the main paper, we present additional quantitative comparisons on Voxceleb2 [3] and LRW [4]. The comparison results outlined in Tab. 1 demonstrate that our method outperforms state-of-the-art approaches in both audio-driven (EDTalk-A) and video-driven (EDTalk-V) scenarios across various metrics. We offer a plausible explanation for the superior $\text{Sync}_{\text{conf}}$ achieved by Wav2Lip [14] in the main paper. IP-LAP [31] merely alters the mouth shape of the source image while maintaining the same head pose and expression, hence achieving a higher PSNR score. PD-FGC [21] attains superior M-LMD performance by training on Voxceleb2, a dataset comprising over **1 million** utterances from **6,112** celebrities, totaling **2400 hours** of data, which is **hundreds of times** larger than our dataset (**15.8 hours**). Nevertheless, we still outperform PD-FGC in terms of F-LMD. SadTalker [30] encounters challenges in processing even one second of audio, leading to the failure to generate talking face videos on the LRW dataset, where all videos are one second in duration.

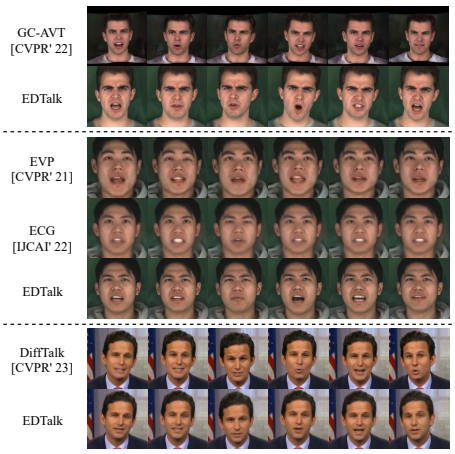

Fig. 3: Comparison results with SOTA methods that have not released their codes and pretrained models.

More qualitative results. In addition to the state-of-the-art (SOTA) methods discussed in the main paper, we extend our comparative analysis to include both emotion-agnostic talking face generation methods: MakeItTalk [33], Wav2Lip [14], Audio2Head [25], AVCT [26], and PC-AVS [32], as well as emotional talking face generation methods: StyleTalk [12] and EMMN [20]. The comprehensive qualitative results can be found in Fig. 2, serving as a supplement to the previously presented data in Fig. 4 of the main paper. We further conduct the comparison experiments with several SOTA talking face generation methods, including: GC-AVT [11], EVP [10], ECG [19] and DiffTalk [17]. However, due to the unavailability of codes and pretrained models for these methods (except EVP), we can only extract video clips from the provided demo videos for comparison. The results are demonstrated in Fig. 3. Specifically, EVP and ECG

are emotional talking face generation methods that utilize one-hot labels for emotional guidance, with EVP being a person-specific model and ECG being a one-shot method. Our method outperforms these methods in terms of emotional expression, while the teeth generated by ECG contribute to slightly unrealistic results. GC-AVT aims to mimic emotional expressions and generate accurate lip motions synchronized with input speech, resembling the setting of our EDTalk. However, compared to EDTalk, GC-AVT struggles to preserve reference identity, resulting in significant identity loss. DiffTalk is hindered by severe mouth jitter, which is more evident in the Supplementary Video.

User Study. We conduct a user study to evaluate our method for *human likeness* test. We generate 10 videos for each method and invite 20 participants (10 males, 10 females) to score from 1 (worst) to 5 (best) in terms of lip-synchronization, realness, and emotion classification. The average scores reported in Tab. 2 demonstrate that our method achieves the best performance in all aspects.

Metric/Method	TalkLip	IP-LAP	EAMM	EAT	EDTalk	GT
Lip-sync	3.31	3.42	3.49	3.85	**4.13**	4.74
Realness	3.14	3.13	3.26	3.75	**4.92**	4.81
Acc_{emo} (%)	19.7	17.6	44.3	59.7	**64.5**	75.6

Table 2: User study results.

2.2 More Comparison with SOTA Face Reenactment Methods

Qualitative results. We perform a comparative analysis with state-of-the-art face reenactment methods, including PIRenderer [15], OSFV [27], LIA [28], DaGAN [8], MCNET [7], StyleHEAT [29], and VPGC [24], where VPGC is a person-specific model. Given that the compared methods are not specifically trained on emotional datasets, we conduct comparisons using videos with and without emotion, the results of which are presented in the Supplementary Video (4:07-4:50). Our method demonstrates superior performance in terms of face reenactment.

Method/Metric	PSNR↑	SSIM↑	LPIPS↓	$\mathcal{L}_1$ ↓	AKD↓	AED↓
PIRenderer [15]	22.13	0.72	0.22	0.053	2.24	0.032
OSFV [27]	23.29	0.74	0.17	0.037	1.83	0.025
LIA [28]	24.75	0.77	0.16	0.036	1.88	0.019
DaGAN [8]	23.21	0.74	0.16	0.041	1.93	0.023
MCNET [7]	21.74	0.69	0.26	0.057	2.05	0.037
StyleHEAT [29]	22.15	0.65	0.25	0.075	2.95	0.045
VPGC [24]	-	-	-	-	-	-
EDTalk	**26.5**	**0.85**	**0.13**	**0.031**	**1.74**	**0.017**

Table 3: The quantitative results compared with SOTA face reenactment methods on HDTF dataset.

Quantitative results. We additionally offer extensive quantitative comparisons regarding: (1) Generated video quality assessed through PSNR and SSIM. (2) Reconstruction faithfulness evaluated using LPIPS and $\mathcal{L}_1$ norms. (3) Semantic consistency measured by average keypoint distance (AKD) and average Euclidean distance (AED). The quantitative results on the HDTF dataset are outlined in Tab. 3, showcasing the superior performance of our EDTalk method.

Note that since VPGC is a person-specific model, it cannot be generalized on identities in HDTF dataset.

2.3 Robustness

Loss functions. We further explored the effects of different loss functions on the MEAD dataset. The results in Table 5 indicate that $\mathcal{L}_{\text{fea}}$ and $\mathcal{L}_{\text{self}}$ contribute to more disentangled spaces, while $\mathcal{L}^m_{\text{rec}}$ and $\mathcal{L}^m_{\text{sync}}$ lead to more accurate lip synchronization. Notably, the **Full Model** shows a reduction in $\text{Sync}_{\text{conf}}$ compared to *only lip* and *lip+pose*, suggesting a trade-off between lip-sync accuracy and emotion performance. In this work, we sacrifice a slight lip-sync accuracy to enhance expression.

Method/Metric	PSNR↑	SSIM↑	M/F-LMD↓	FID↓	$\text{Sync}_{\text{conf}}$ ↑	Acc_{emo} ↑
w/o $\mathcal{L}_{fea}$	21.134	0.713	1.914/1.625	28.053	5.601	54.34
w/o $\mathcal{L}_{self}$	20.913	0.707	1.815/1.629	29.314	5.030	44.23
w/o $\mathcal{L}^m_{rec}$	**21.955**	**0.744**	1.666/1.397	18.528	5.447	67.19
w/o $\mathcal{L}^m_{sync}$	21.524	0.728	1.626/1.349	17.844	4.007	61.29
w/o Orthogonal	21.429	0.711	1.687/1.320	17.820	4.398	38.71
w/o Bank	20.302	0.660	2.137/1.711	26.842	2.316	9.677
w/o EEM	20.731	0.673	2.131/1.927	27.135	7.326	49.367
only lip	19.799	0.639	1.767/1.920	31.918	8.291	15.13
lip+pose	21.519	0.695	1.645/1.378	19.571	**8.474**	16.75
Full Model	21.628	0.722	**1.537**/**1.290**	**17.698**	8.115	**67.32**

Table 4: Ablation study results.

2.4 Robustness

Our method demonstrates robustness across out-of-domain portraits, encompassing real human subjects, paintings, sculptures, and images generated by Stable Diffusion [16]. Moreover, our approach exhibits generalizability to various audio inputs, including songs, diverse languages (English, French, German, Italian, Japanese, Korean, Spanish, Chinese), and noisy audio. Please refer to the Supplementary Video (5:40-8:40) for the better visualization of these results.

2.5 Expression Manipulation

We accomplish expression manipulation by interpolating between expression weights W^e of the expression bank B^e, which are extracted from any two distinct expression reference clips, using the following equation:

$$W^e = \alpha W^e_1 + (1-\alpha) W^e_2, \tag{3}$$

where W^e_1 and W^e_2 represent expression weights extracted from two emotional clips, while α denotes the interpolation weight. Fig. 4 illustrates an example of expression manipulation generated by our EDTalk. In this example, we successfully transition from $Expression1$ to $Expression2$ by varying the interpolation weight α. This demonstrates the effectiveness of our ELN module in accurately capturing the expression of the provided clip, as discussed in the main paper.

2.6 Probabilistic Pose Generation

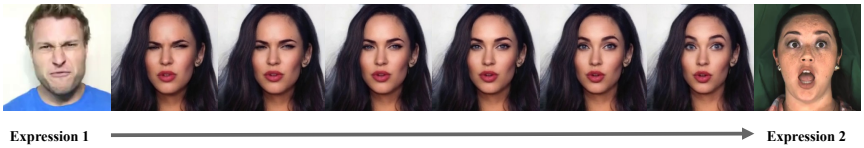

Fig. 4: The results of expression manipulation.

Thanks to the distribution p_Z modeled by the Audio2Pose module, we are able to sample diverse and realistic head poses from it. As shown in Fig. 5, by passing the same inputs through our EDTalk, our method synthesizes various yet natural head motions while preserving the expression and mouth shape unchanged.

Fig. 5: The results of generated head poses.

2.7 Semantically-Aware Expression Generation

We input two transcripts into a Text-To-Speech (TTS) system to synthesize two audio clips. These audios, along with their respective transcripts, are then fed into our Audio-to-Motion module to generate talking face videos. The results of semantically-aware expression generation are depicted in Fig. 6, showcasing our method's ability to accurately generate expressions corresponding to the transcripts (left: happy; right: sad). Additionally, in the Supplementary Video, we provide further results where expressions are inferred directly from audio.

2.8 Motion Direction Controlled by Base

We initially present the results showcasing individual control over mouth shape, head pose, and emotional expression in Fig. 7. Specifically, by feeding our EDTalk

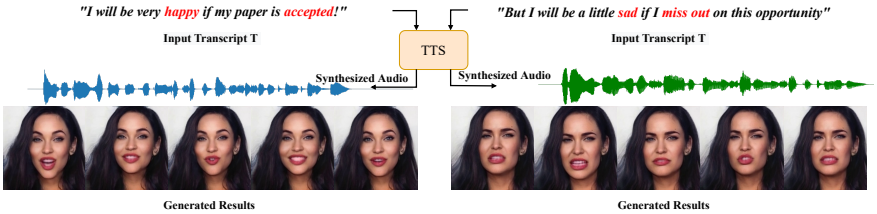

Fig. 6: The results of semantically-aware expression generation.

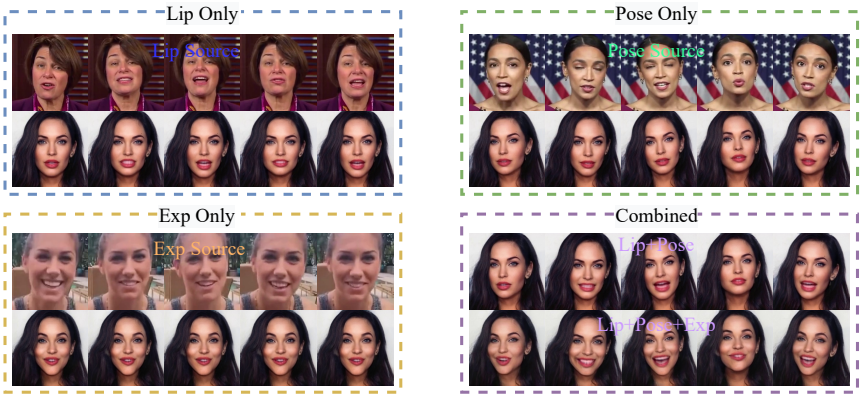

Fig. 7: The results of individual control over mouth shape, head pose, emotional expression and combined facial dynamics.

Fig. 8: Motion direction controlled by each base.

with an identity source and various driving sources (first row of each part), our method generates corresponding disentangled outcomes in the second row. Subsequently, we integrate these individual facial motions into full emotional talking head videos with synchronized lip movements, head gestures, and emotional expressions. It's worth noting that our method facilitates the combination of any two facial parts, such as 'expression+lip', 'expression+pose', etc. An example of 'lip+pose' is shown in the first row in the lower right corner of Fig. 7. Additionally, we provide comparisons with state-of-the-art facial disentanglement methods like PD-FGC [21] and DPE [13] in terms of facial disentanglement performance and computational efficiency. For further details, please refer to the Supplementary Video (4:50-5:12).

We are also intrigued by understanding how each base in the banks influences motion direction. Consequently, we manipulate only a specific base b^*_i

Method	Mouth Bank B^m					Expression Bank B^e				
	PSNR↑	SSIM↑	M/F-LMD↓	$\text{Sync}_{\text{conf}}$ ↑	Acc_{emo} ↑	PSNR↑	SSIM↑	M/F-LMD↓	$\text{Sync}_{\text{conf}}$ ↑	Acc_{emo} ↑
5	20.39	0.69	2.02/1.67	6.35	63.53	21.54	0.70	1.60/1.35	8.27	53.26
10	21.45	**0.72**	1.65/1.33	7.89	65.74	**21.63**	**0.72**	**1.54/1.29**	**8.12**	**67.32**
20	**21.63**	**0.72**	**1.54/1.29**	**8.12**	67.32	21.37	**0.72**	1.64/1.46	8.23	61.34
40	20.79	0.71	1.65/1.48	7.62	63.12	21.41	0.71	1.68/1.42	8.16	59.65

Table 5: Ablation study on the number of base.

and repeat the setup. The results, as depicted in Fig. 8, indicate that the bases hold semantic significance for fundamental visual transformations such as mouth opening/closing, head rotation, and happiness/sadness/anger.

2.9 Ablation Study

Bank size. In this section, we perform a series of experiments on the MEAD dataset to explore the impact of base number selection on final performance. Specifically, we vary the base number of the Mouth Bank B^m and Expression Bank B^e across values of 5, 10, 20, and 40, respectively. The quantitative results are provided in Tab. 5, where we observe the best performance when utilizing 20 bases in B^m and 10 bases in B^e.

3 Discussion

3.1 Novelty

Our approach is efficient thanks to the constraints we impose on the latent spaces (requirement (a), (b)). Based on these requirements, we propose a simple and easy-to-implement framework and training strategy. This does not require large amounts of training time, training data, and computational resources. However, it does not indicate a lack of innovation in our approach. Quite the contrary, in an age where computational power reigns, our aim is to propose an efficient strategy that attains state-of-the-art performance with minimal computational resources, eschewing complex network architectures or training gimmicks. We aspire for our method to offer encouragement and insight to researchers operating within resource-constrained environments, presented in a simple and elegant manner!

3.2 Why Different Disentanglement Strategies?

The design is based on: **(1) Physical nature.** On one hand, both mouth shape and emotional expression are performed on face region with **mutual influence**, while the head pose is **independent** of them, manifested as whole head rotation and translation. It makes pose disentanglement **easier** than mouth and expression. On the other hand, in facial dynamics, mouth moves more **frequently** than emotional expressions, which makes mouth feature extraction **easier** than

emotional expression. These physical natures motivate us to simplify the whole decoupling process by first decoupling head pose and mouth shape (a relatively **simpler** task) and then decoupling emotional expression (a more **complex** task). **(2) Dataset.** Current datasets can be broadly divided into two types: those comprising neutral expressions with diverse poses (e.g., HDTF) and those containing diverse expressions with subtle poses (e.g., MEAD). Such datasets also motivate us to leverage the former (where there are no **distractions** from **expressions**) for pose/mouth disentanglement and the latter for expression disentanglement.

3.3 Potential Worries about Mouth-pose Decouple Module

'Pose' or 'Non-Mouth'? Since we only replace the mouth regions of the data during training mouth-pose decouple module, the decoupled 'pose' space in this stage actually refers to the 'non-mouth' region, including expression and head pose. To mitigate the influence of expression on this pose space, we exclusively train with an expression-agnostic dataset, where all images maintain a neutral expression. As a result, the mouth-pose decouple module in this stage solely focuses on the head pose and lacks the capability to model emotive expression. Therefore, we refer to it as 'pose' instead of 'non-mouth'. This hypothesis was further validated in our experiments (Fig. 7 and Fig. 8); even when emotional videos are inputted, the PLN module solely extracts head pose without incorporating emotional expression.

Color Artifact caused by replacing mouth. We notice that there exist some color artifacts in synthesized images (pointed by red arrows in Fig. 9). However, we argue that these artifacts do not significantly impact performance and provide a detailed analysis to support this claim. **(1)** Our Encoder E and Generator G are pretrained in a similar setting as LIA [28], using a dataset collected from various sources with diverse identities, backgrounds, and motions. This diversity results in richness and colorfulness in each frame, making the Encoder E robust to different input images. We have verified this robustness in our experiments (see Sec. 2.4). Therefore, despite the presence of artifacts, the Encoder E can effectively process synthetic images. **(2)** During the training process, we employ not only cross-reconstruction but also self-reconstruction loss ($\mathcal{L}_{self}$) on images without mouth replacement. This loss makes the training data contain not only synthesized images but also a large number of unmodified images, thereby preventing performance degradation. We have also confirmed the contribution of self-reconstruction through our ablation study.

Fig. 9: Examples of synthesized images. $I_{p_A}^{m_B}$ refers to image A with the mouth of B, and vice versa.

Comparison Protocol. One might raise concerns regarding the evaluation datasets, as both MEAD and HDTF datasets used for evaluation are also the datasets on which the model is trained. Moreover, several prior works used for comparison haven't been trained on the HDTF dataset. For instance, PD-FGC isn't trained on the HDTF dataset, raising questions about the fairness of such comparisons. We provide several explanations to address these concerns: **(1)** To maintain consistency with previous works, we adhere to the comparison protocol established by them [9, 12]. Specifically, both MEAD and HDTF datasets contain a mix of 43 available speakers and over 300 speakers. We randomly allocate 4 and 60 speakers for testing and the remainder for training. This ensures that the test set comprises identities unseen during training, thereby ensuring a fair comparison. **(2)** While some works, such as PD-FGC, aren't trained on the HDTF or MEAD datasets, they utilized the Voxceleb2 dataset, which includes **over 1 million** utterances from **6,112** celebrities. This dataset size is hundreds of times larger than ours, ensuring that they have ample data for training. **(3)** Additionally, we conduct comparisons on the LRW and Voxceleb2 datasets, which are not utilized by our method. The results presented in Tab. 1 reaffirm the superiority of our approach, providing further validation of the performance.

Limitation While our current work has made significant strides, it also possesses certain limitations. Firstly, due to the low resolution of the training data, our approach is constrained to generating videos with a resolution of 256×256. Consequently, the blurred teeth in the generated results may diminish their realism. Secondly, our method currently overlooks the influence of emotion on head pose, which represents a meaningful yet unexplored task. Unfortunately, the existing emotional MEAD dataset [23] maintains consistent head poses across emotions, making it challenging to model the impact of emotion on pose. However, once relevant datasets become available, our approach can readily be extended to incorporate the influence of emotion on head pose by introducing emotion labels e as an additional conditioning factor, as depicted in Eq. (13): $\hat{W}^p = \varphi_p(z, f_t^a, e)$.

Ethical considerations. Our approach is geared towards generating talking face animations with individual facial control, which holds promise for various applications such as entertainment and filmmaking. However, there is a potential for malicious misuse of this technology on social media platforms, leading to negative societal implications. Despite significant advancements in deepfake detection research [1, 2, 6, 34], there is still room for improvement in detection accuracy, particularly with the availability of more diverse and comprehensive datasets. In this regard, we are pleased to offer our talking face results, which can contribute to enhancing detection algorithms to better handle increasingly sophisticated scenarios.

References

1. Arandjelovic, R., Zisserman, A.: Look, listen and learn. In: Proceedings of the IEEE international conference on computer vision. pp. 609–617 (2017)

2. Chai, L., Bau, D., Lim, S.N., Isola, P.: What makes fake images detectable? understanding properties that generalize. Lecture Notes in Computer Science (2020)
3. Chung, J.S., Nagrani, A., Zisserman, A.: Voxceleb2: Deep speaker recognition. arXiv preprint arXiv:1806.05622 (2018)
4. Chung, J.S., Zisserman, A.: Lip reading in the wild. In: Computer Vision–ACCV 2016: 13th Asian Conference on Computer Vision, Taipei, Taiwan, November 20-24, 2016, Revised Selected Papers, Part II 13. pp. 87–103. Springer (2017)
5. Gan, Y., Yang, Z., Yue, X., Sun, L., Yang, Y.: Efficient emotional adaptation for audio-driven talking-head generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 22634–22645 (2023)
6. Guera, D., Delp, E.J.: Deepfake video detection using recurrent neural networks. Advanced Video and Signal Based Surveillance (2018)
7. Hong, F.T., Xu, D.: Implicit identity representation conditioned memory compensation network for talking head video generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 23062–23072 (2023)
8. Hong, F.T., Zhang, L., Shen, L., Xu, D.: Depth-aware generative adversarial network for talking head video generation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3397–3406 (2022)
9. Ji, X., Zhou, H., Wang, K., Wu, Q., Wu, W., Xu, F., Cao, X.: Eamm: One-shot emotional talking face via audio-based emotion-aware motion model. In: ACM SIGGRAPH 2022 Conference Proceedings. pp. 1–10 (2022)
10. Ji, X., Zhou, H., Wang, K., Wu, W., Loy, C.C., Cao, X., Xu, F.: Audio-driven emotional video portraits. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 14080–14089 (2021)
11. Liang, B., Pan, Y., Guo, Z., Zhou, H., Hong, Z., Han, X., Han, J., Liu, J., Ding, E., Wang, J.: Expressive talking head generation with granular audio-visual control. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3387–3396 (2022)
12. Ma, Y., Wang, S., Hu, Z., Fan, C., Lv, T., Ding, Y., Deng, Z., Yu, X.: Styletalk: One-shot talking head generation with controllable speaking styles. arXiv preprint arXiv:2301.01081 (2023)
13. Pang, Y., Zhang, Y., Quan, W., Fan, Y., Cun, X., Shan, Y., Yan, D.m.: Dpe: Disentanglement of pose and expression for general video portrait editing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 427–436 (2023)
14. Prajwal, K., Mukhopadhyay, R., Namboodiri, V.P., Jawahar, C.: A lip sync expert is all you need for speech to lip generation in the wild. In: Proceedings of the 28th ACM International Conference on Multimedia. pp. 484–492 (2020)
15. Ren, Y., Li, G., Chen, Y., Li, T.H., Liu, S.: Pirenderer: Controllable portrait image generation via semantic neural rendering. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13759–13768 (2021)
16. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models (2021)
17. Shen, S., Zhao, W., Meng, Z., Li, W., Zhu, Z., Zhou, J., Lu, J.: Difftalk: Crafting diffusion models for generalized audio-driven portraits animation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1982–1991 (2023)
18. Siarohin, A., Lathuilière, S., Tulyakov, S., Ricci, E., Sebe, N.: First order motion model for image animation. Advances in neural information processing systems **32** (2019)

19. Sinha, S., Biswas, S., Yadav, R., Bhowmick, B.: Emotion-controllable generalized talking face generation
20. Tan, S., Ji, B., Pan, Y.: Emmn: Emotional motion memory network for audio-driven emotional talking face generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 22146–22156 (2023)
21. Wang, D., Deng, Y., Yin, Z., Shum, H.Y., Wang, B.: Progressive disentangled representation learning for fine-grained controllable talking head synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 17979–17989 (2023)
22. Wang, J., Qian, X., Zhang, M., Tan, R.T., Li, H.: Seeing what you said: Talking face generation guided by a lip reading expert. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14653–14662 (2023)
23. Wang, K., Wu, Q., Song, L., Yang, Z., Wu, W., Qian, C., He, R., Qiao, Y., Loy, C.C.: Mead: A large-scale audio-visual dataset for emotional talking-face generation. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXI. pp. 700–717. Springer (2020)
24. Wang, K., Zhou, H., Wu, Q., Tang, J., Xu, Z., Liang, B., Hu, T., Ding, E., Liu, J., Liu, Z., et al.: Efficient video portrait reenactment via grid-based codebook. In: ACM SIGGRAPH 2023 Conference Proceedings. pp. 1–9 (2023)
25. Wang, S., Li, L., Ding, Y., Fan, C., Yu, X.: Audio2head: Audio-driven one-shot talking-head generation with natural head motion. In: International Joint Conference on Artificial Intelligence. IJCAI (2021)
26. Wang, S., Li, L., Ding, Y., Yu, X.: One-shot talking face generation from single-speaker audio-visual correlation learning. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 36, pp. 2531–2539 (2022)
27. Wang, T.C., Mallya, A., Liu, M.Y.: One-shot free-view neural talking-head synthesis for video conferencing. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10039–10049 (2021)
28. Wang, Y., Yang, D., Bremond, F., Dantcheva, A.: Latent image animator: Learning to animate images via latent space navigation. In: International Conference on Learning Representations (2021)
29. Yin, F., Zhang, Y., Cun, X., Cao, M., Fan, Y., Wang, X., Bai, Q., Wu, B., Wang, J., Yang, Y.: Styleheat: One-shot high-resolution editable talking face generation via pre-trained stylegan. In: European conference on computer vision. pp. 85–101. Springer (2022)
30. Zhang, W., Cun, X., Wang, X., Zhang, Y., Shen, X., Guo, Y., Shan, Y., Wang, F.: Sadtalker: Learning realistic 3d motion coefficients for stylized audio-driven single image talking face animation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8652–8661 (2023)
31. Zhong, W., Fang, C., Cai, Y., Wei, P., Zhao, G., Lin, L., Li, G.: Identity-preserving talking face generation with landmark and appearance priors. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9729–9738 (2023)
32. Zhou, H., Sun, Y., Wu, W., Loy, C.C., Wang, X., Liu, Z.: Pose-controllable talking face generation by implicitly modularized audio-visual representation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4176–4186 (2021)
33. Zhou, Y., Han, X., Shechtman, E., Echevarria, J., Kalogerakis, E., Li, D.: Makelttalk: speaker-aware talking-head animation. ACM Transactions On Graphics (TOG) **39**(6), 1–15 (2020)

34. Zhou, Y., Lim, S.N.: Joint audio-visual deepfake detection. International Conference on Computer Vision (2021)