FaceChain-ImagineID: Freely Crafting High-Fidelity Diverse Talking Faces from Disentangled Audio

Chao Xu[1*] Yang Liu[1*] Jiazheng Xing[2] Weida Wang[2] Mingze Sun[2]
Jun Dan[2] Tianxin Huang[3] Siyuan Li[2] Zhi-Qi Cheng[4] Ying Tai[5] Baigui Sun[1†]
[1]Alibaba Group [2]FaceChain Community [3]National University of Singapore
[4]Carnegie Mellon University [5]Nanjing University
{xc264362, ly261666, baigui.sbg}@alibaba-inc.com

We supplement the following contents, which are not presented in the paper due to space limitations:

- Preliminaries
- More discussions about PAD
- Details of model architecture
- Additional experiments
- Limitations
- Societal impacts

0.1. Preliminaries

3D Morphable Models. Follow the previous work [1] that receives an input face $\boldsymbol{I}$ and estimates the 3DMMs coefficient $\boldsymbol{\phi}$, including identity $\boldsymbol{\alpha} \in \mathbb{R}^{80}$, expression $\boldsymbol{\beta} \in \mathbb{R}^{64}$, texture $\boldsymbol{\delta} \in \mathbb{R}^{80}$, illumination $\boldsymbol{\gamma} \in \mathbb{R}^{27}$, and pose $\boldsymbol{p} \in \mathbb{R}^{6}$. Formally:

$$\boldsymbol{\phi} = \{\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\delta}, \boldsymbol{\gamma}, \boldsymbol{p}\}\,. \tag{1}$$

With 3DMM $\boldsymbol{\phi}$, the 3D shape $\mathbf{S}$ and albedo texture $\mathbf{T}$ could be parameterized as:

$$\begin{aligned} \mathbf{S} &= \bar{\mathbf{S}} + \mathbf{B}_{id}\boldsymbol{\alpha} + \mathbf{B}_{exp}\boldsymbol{\beta}, \\ \mathbf{T} &= \bar{\mathbf{T}} + \mathbf{B}_{t}\boldsymbol{\delta}, \end{aligned} \tag{2}$$

where $\bar{\mathbf{S}}$ and $\bar{\mathbf{T}}$ denote the mean face shape and albedo texture. $\mathbf{B}_{id}$, $\mathbf{B}_{exp}$, and $\mathbf{B}_{t}$ are the bases of identity, expression, and texture computed via PCA. We project the reconstructed 3D face onto the 2D image plane with a differentiable renderer $\boldsymbol{\mathcal{R}}$ according to its illumination $\boldsymbol{\gamma}$ and $\boldsymbol{p}$:

$$\boldsymbol{I}_{rd} = \boldsymbol{\mathcal{R}}(\mathbf{S}, \mathbf{T}, \boldsymbol{\gamma}, \boldsymbol{p}). \tag{3}$$

Please refer to D3DFR [1] and its officially released code[1] for more details.

Latent Diffusion Models. LDMs [7] first employ a encoder $\boldsymbol{\mathcal{E}}$ that project an image $\boldsymbol{I}$ into a latent $\boldsymbol{z} = \boldsymbol{\mathcal{E}}(\boldsymbol{I})$, which can be reconstructed back to the image $\boldsymbol{I} \approx \boldsymbol{\mathcal{D}}(\boldsymbol{z})$ by decoder $\boldsymbol{\mathcal{D}}$. A U-Net ϵ_θ containing self-attention and cross-attention is followed to remove the noise with the objectives:

$$\min_{\theta} E_{\boldsymbol{z}_0, \boldsymbol{\varepsilon} \sim N(0,1), t \sim \text{ Uniform } (1,T)} \left\|\boldsymbol{\varepsilon} - \boldsymbol{\varepsilon}_\theta\left(\boldsymbol{z}_t, t, \boldsymbol{P}\right)\right\|_2^2, \tag{4}$$

where $\boldsymbol{P}$ is the embedding of the conditional text prompt and $\boldsymbol{z}_t$ is a noisy sample of $\boldsymbol{z}_0$ at timestep t.

[1]https://github.com/sicxu/Deep3DFaceRecon_pytorch/tree/master

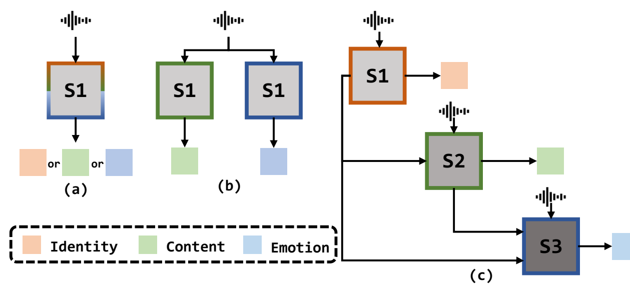

Figure 1. The comparison of recent works and our PAD during inference. S means stage.

0.2. More Discussions about PAD

The Disentanglement Order. We adhere to the principle of prioritizing the disentanglement of easier and cleaner elements first. Concretely, identity provides the basic facial bone structure and the position of facial features, including the mouth's position and shape. Based on this foundation, as shown in Fig. 5(b) of the main paper, the content primarily involves the *movement around the lip* while the upper face remains almost fixed (cols. 4-5). Thus dubbing methods, *e.g.*, Wav2Lip [6], IP-LAP [11] only edit the bottom face. However, the emotion involves not only the local lip movement but also *global facial deformation*, *i.e.*, the same spoken content exhibits distinct variations in mouth, eye, and eyebrow across different emotions (cols. 1-4). Therefore, we adopt the identity → content → emotion paradigm.

The Superior Properties. The proposed PAD offers new insights into effectively disentangling multiple highly cou-

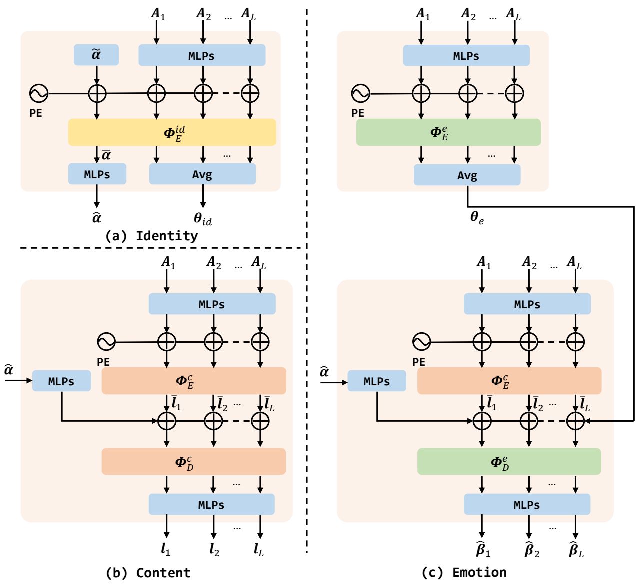

Figure 2. (a) Architecture of Identity Disentanglement. (b) Architecture of Content Disentanglement. (c) Architecture of Emotion Disentanglement.

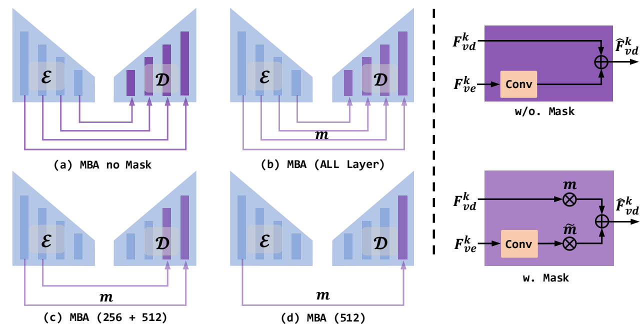

Figure 3. Architecture of four MBA variants.

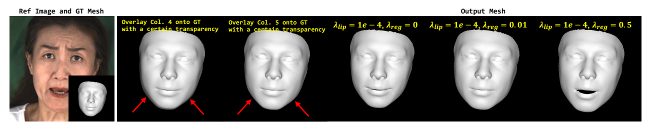

Figure 4. Ablation study of regularization loss in Content Disentanglement. The GT mesh is rendered from the disentangled shape and zero-initialization expression (closed mouth).

Method	Audio-Visual Sync. ↑	Emo Acc. ↑	Video Quality ↑
Wav2Lip	4.16	3.52	3.29
PC-AVS	4.11	3.70	3.26
EAMM	3.63	3.77	2.89
SadTalker	4.09	3.75	3.97
Ours	**4.21**	**4.05**	**4.17**

Table 1. The results of user study on test set.

pled factors within a unified framework, especially in cases where there is a lack of sufficient annotated data for large-scale supervised learning to directly separate all elements. 1) *Differentiation and highlighting*: As shown in Fig. 1, (a) most works [3, 4, 9] only extract single cue from audio. (b) Some studies [2, 5] build pseudo pairs to disentangle two elements (content and emotion) under cross-reconstruction, yet such preprocessing inevitably introduces errors. In contrast, (c) our method covers *three* factors (identity, content, emotion) and considers their intricate relationships, tailoring a progressive approach to gradually decouple each of them. 2) *Strengths*: PAD introduces accurate disentangled 3D facial prior [1], and each stage is only responsible for a specific factor, thus reducing the difficulty and improving the performance.

0.3. Details of Model Architecture

The Detailed Architecture of PAD. In Fig. 2, identity encoder $\mathbf{\Phi}_E^{id}$, content encoder $\mathbf{\Phi}_E^{c}$, emotion encoder $\mathbf{\Phi}_E^{e}$, content decoder $\mathbf{\Phi}_D^{c}$, and emotion decoder $\mathbf{\Phi}_D^{e}$ are built upon Transformer [8] networks with 4 layers, 8 heads, and 512 latent feature dimensions. MLPs consist of two Linear layers. The whole process is described in the main paper.

Four Variants of MBA in Ablation Study. We show the architectures of four MBA variants used in Sec. 4.4 of the main paper. Fig 3(a) is the MBA without mask-guided blending, (b) with mask-guided blending applied on all layers, (c) is applied at 256 and 512 resolutions, (d) is only applied at 512 resolution. We adopt the structure of (d) in our method, which achieves the best visual performance.

0.4. Additional Experiments

User Study. We conduct a user study to evaluate the performance of four talking face generation methods with officially released codes, *i.e.*, Wav2Lip [6], PC-AVS [12], EAMM [3], and SadTalker [10]. We randomly sample 50 videos from the test set and require 20 participants to evaluate the given videos from three dimensions: 1) Audio-visual synchronization; 2) Emotion accuracy; 3) Overall video quality by rating scores from 1 to 5. We average the scores and report the results in Tab 1, illustrating that our method significantly surpassed the other methods.

The Effectiveness of Regularization Loss in Content Disentanglement. This loss constraints the linguistic features $\boldsymbol{l}$ into the 3DMM domain for more stable and faster training. To verify its effectiveness, we supplement an ablation study under the same training setting (*e.g.*, training steps, learning rate). An angry but silent face is shown in Fig, whose mouth is expected to be *closed* during the content disentanglement. The face shape is disturbed without $\mathcal{L}_{reg}$ (cols. 2, 3 depict the difference, the contour becomes sharper w/o it), and satisfactory results are achieved when the weight is set to 0.01 (col. 5). Thus, the proper weight of $\mathcal{L}_{reg}$ benefit this stage.

0.5. Limitations

Our method generates a frame as an iterative denoising process, which needs more time compared with most GAN-based approaches. Besides, even though we have implemented several conditions to achieve coherent frame generation, the synthesized video still exhibits slight flickering, thus appearing to lack temporal consistency. These are also common problems of LDM-based works. Although we use face mesh and identity face to provide the appearance and structure guidance of mouth areas, but encounter challenges in accurately representing the teeth, resulting in inconsistent temporal changes around this region, which also shows artifacts and an unrealistic appearance like most recent methods.

0.6. Societal Impacts

The advancement of talking face generation has garnered significant attention and is applied for various ethical and legitimate purposes, including in films and virtual reality. However, like any technology, it has the potential for both positive and negative applications. We maintain a zero-tolerance policy against any unethical use of our work and actively discourage such misuse.

References

[1] Yu Deng, Jiaolong Yang, Sicheng Xu, Dong Chen, Yunde Jia, and Xin Tong. Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 0–0, 2019. 1, 2

[2] Xinya Ji, Hang Zhou, Kaisiyuan Wang, Wayne Wu, Chen Change Loy, Xun Cao, and Feng Xu. Audio-driven emotional video portraits. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14080–14089, 2021. 2

[3] Xinya Ji, Hang Zhou, Kaisiyuan Wang, Qianyi Wu, Wayne Wu, Feng Xu, and Xun Cao. Eamm: One-shot emotional talking face via audio-based emotion-aware motion model. In *ACM SIGGRAPH 2022 Conference Proceedings*, pages 1–10, 2022. 2

[4] Youxin Pang, Yong Zhang, Weize Quan, Yanbo Fan, Xiaodong Cun, Ying Shan, and Dong-ming Yan. Dpe: Disentanglement of pose and expression for general video portrait editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 427–436, 2023. 2

[5] Ziqiao Peng, Haoyu Wu, Zhenbo Song, Hao Xu, Xiangyu Zhu, Jun He, Hongyan Liu, and Zhaoxin Fan. Emotalk: Speech-driven emotional disentanglement for 3d face animation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20687–20697, 2023. 2

[6] KR Prajwal, Rudrabha Mukhopadhyay, Vinay P Namboodiri, and CV Jawahar. 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*, pages 484–492, 2020. 1, 2

[7] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 1

[8] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2

[9] Yandong Wen, Bhiksha Raj, and Rita Singh. Face reconstruction from voice using generative adversarial networks. *Advances in neural information processing systems*, 32, 2019. 2

[10] Wenxuan Zhang, Xiaodong Cun, Xuan Wang, Yong Zhang, Xi Shen, Yu Guo, Ying Shan, and Fei Wang. 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*, pages 8652–8661, 2023. 2

[11] Weizhi Zhong, Chaowei Fang, Yinqi Cai, Pengxu Wei, Gangming Zhao, Liang Lin, and Guanbin Li. Identity-preserving talking face generation with landmark and appearance priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2023. 1

[12] Hang Zhou, Yasheng Sun, Wayne Wu, Chen Change Loy, Xiaogang Wang, and Ziwei Liu. Pose-controllable talking face generation by implicitly modularized audio-visual representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4176–4186, 2021. 2