

Supplementary Material of MobilePortrait

1. Discussion of Potential Negative Impact

This paper proposes a method that can generate a video from a single facial image based on specified video or audio, which carries a potential risk of being used for nefarious purposes and leading to negative social impacts, as with most AIGC-related methods. To avoid these potential negative effects, we consider measures beyond regulation: 1. Restricting usage to non-real human images only; 2. Real-human animation requires liveness and identity verification; 3. Limiting content to preset templates to prevent malicious use; 4. Adding visible labels to AIGC content; 5. Embedding watermarks for traceability.

2. Visualization Results of MobilePortrait

Supplementary videos demonstrate our method’s effectiveness and compare it with others using visuals from the main text, with some showing solid-color padding due to cropping. These videos feature both video-driven and audio-driven results, as well as failure cases. As explained in Table 1, all videos were uniformly resized to 256px for fair comparison, since some methods only output at this resolution. Additionally, we provide higher-resolution 512 results for MobilePortrait. Based on our audio-driven video results, our method achieves results that are comparable to other methods. In fact, it is somewhat unfair to use the SyncNet proposed in Wav2Lip [3] to evaluate our method, as the comparative methods, including SadTalker [6] and Real3D [5], utilize the SyncNet or Wav2Lip model for auxiliary training during their training of audio-to-motion modules.

Table 1. The resolution of the output from the current methods.

Method	FaceV2V	TPS	MCNet	Real3D	Ours
FLOPs	629G	140G	200G	610G	16G
Resolution	256	256	256	512	512

3. Model Architecture

This section describes the model structures for key components of MobilePortrait, drawing upon architectures

Table 2. UNet Block Modification

UNet	Down&UpBlock	MiddleBlock
Original	3x3C	3x3C-3x3C
Modified	1x1C-3x3C-1x1C	1x1C-3x3DepC-1x1C

from prior work. Our Dense Motion Network, following FOMM [4], TPS [7] and MCNet [2], adopts a U-Net-like configuration with five down blocks featuring 3×3 convolutions and batch normalization, which reduce spatial dimensions by half at each level, with a progression of channels from 128 to 1024. The up blocks are structured inversely to the down blocks. In the synthesis network, we maintain the simple U-Net architecture without incorporating the additional components proposed by previous works [2, 7]. As shown in the Table 2 (excluding ReLU and BN, ‘C’ for convolution), we replaced the Original ConvBlock with a bottleneck structure to reduce FLOPs, assigning 96 output channels to the first six down blocks and 128 to the latter six, with a 2x downsampling every three blocks, mirrored in the up blocks. The middle block was replaced with a 3x3 depthwise convolution bottleneck, repeated 8 times with 192 output channels and 4x channel multiplier. This change resulted in the UNet (8 GFLOPs and 8M parameters) for our 16G model, with remaining 3x3 convolutions replaced by depthwise convolutions in smaller models.

For other parts, ResNet18 [1] is employed as the backbone for the neural keypoint detector. A LSTM (3 layers, 1024 channels, 0.6 GFLOPs) are used for audio-to-motion module.

4. Limitations of MobilePortrait

Although MobilePortrait, as demonstrated in the main text, maintains good robustness when handling most images and motions, from Figure 1 and Figure 2, we can observe that it still struggles when dealing with extreme angles of motion or styles that differ significantly from the training data. We speculate that this is because, for the image synthesis network, there is a need to inpaint a large amount of content in these scenarios, which often includes patterns that are difficult to learn from the training data, such as large areas of profile faces or cartoon styles. One solution is to address

this issue by increasing the diversity of the training dataset, while another is to rely on robust intermediate representations, such as 3D facial structures. We leave this as future work to be tackled.



Figure 1. **Visualization result.** Blurriness Caused by Mismatched Art Styles



Figure 2. **Visualization result.** The generation of invisible areas in the source image is not good when the turning angle is too large.

5. Extended Experimental Results

Effectiveness of MobilePortrait. We further conducted two experiments based on the state-of-the-art method MCNet: 1) Replace their UNet but keep MemUnit (not feasible for smartphone inference; MCNet’s original UNet requires 50M parameters and 190 GFLOPs, with its MemUnit taking 25M parameters and 65 GFLOPs. Now the UNet is replaced by our modified version with only convolutional layers, but the 65 GFLOPs MemUnits are kept). Same-id reenactment results are in Table 4, with other metrics performing consistently. 2) Further removing MemUnit, we return to a baseline (TPS based model) as already shown in Figure 4 in the main text, with visual comparisons in Figure 5. We observe a significant performance drop without the proposed modules under reduced computational load, indicating that the external knowledge modules are key to high-quality results. Simply reducing computational load by replacing the backbone is not feasible. This conclusion is also supported by ablation studies in the main text, reaffirming our approach’s effectiveness.

Table 3. **Ablation studies of training datasets.**

Method	FID ↓	AKD ↓	APD(C) ↓	AED(C) ↓	CSIM(C) ↑
Full Datasets	29.2	1.30	2.7	0.13	39.2
remove VoxCelebvHQ	32.5	1.43	2.9	0.13	37.7
remove VFHQ	37.1	1.49	3.6	0.13	38.5

Table 4. **Key results compared with MCNet**

Method	FID	AKD
Baseline with MemUnit	78.7	2.46
Ours	29.2	1.30

Table 5. **Ablation studies.** AKD is evaluated under the same-ID setting, while AED and HPD are evaluated under the cross-ID setting.

Method	AED	AKD	HPD
Mixed Keypoint.	0.13	1.30	2.74
NK-Only	0.17	2.62	3.90
FK-Only	0.13	1.61	10.5
No Proposed Loss	0.13	1.45	4.19

Method	AED	AKD	HPD
Ours	0.13	1.30	2.74
Trans. Fusing KP	0.13	3.14	9.5
Conv. Fusing KP	0.15	1.49	3.7
No Residual O.F.	0.14	1.45	6.2

Inp. BG FG Comp.	AED	AKD	HPD
	0.14	1.54	7.3
✓	0.13	1.30	2.74
	0.13	1.52	5.70
✓	0.13	1.47	10.0

#Views	AED	AKD	HPD
0	0.13	2.53	3.21
2	0.13	1.53	3.07
4	0.13	1.30	2.74
8	0.13	1.31	2.1

Training Data. Training data-related experimental results are provided in Table 3. We sequentially removed the Vox and VFHQ datasets until only CelebVHQ remained. Results show performance declines with dataset removal, mainly affecting the FID image quality index, with a smaller impact on motion accuracy. Even with only CelebVHQ, we still outperform the recent method Real3D. Combined with Table 1 in the main text, it can be seen that even with reduced data, our method still matches or exceeds some recent methods with higher FLOPs.

AED Results. Due to space limitations in the main text, we couldn’t include the differences in AED (cross-id evaluation) metrics from the comparative experiments. Here, we supplement this information in Table ?? . Although the differences in AED are smaller compared to AKD and HPD, the proposed method still demonstrates better performance.

Visual Results on Large Poses. We also provide visualized test results for extreme angles in Figure 3. MobilePortrait 16G achieves similar results to larger models and significantly outperforms the baseline.

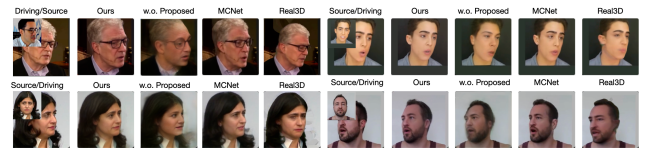


Figure 3. **Visualization result.** Generated frames driven by large pose.

6. Supplementary implementation details

In module training, we used a pretrained ResNet18-based face 106-keypoints detector as the face landmark extractor. The FLOPs statistics account for our end-to-end trained model but do not include the pretrained face keypoint detector. However, the inference time calculation does include the time for the face keypoint detector. The synthesis network, including the appearance feature extractor, is trained in an end-to-end manner. For the pseudo background, dur-

ing training, we use the ground truth (GT) image of each generated frame, inpainted as the pseudo background. This approach helps the model trust the provided background, as it aligns with the training targets. If we were to provide this pseudo background during training, the model would need to learn additional synthesis capabilities due to misalignments, thereby increasing its burden. Instead, during inference, we provide a pseudo background based on the source image.

When testing speed, we recorded single inference times for the CoreML model with a 512x512 output. On the iPhone 14 Pro with the A16 chip, all computations are successfully performed on the NPU. On the iPhone 12 with the A14 chip, only 7

References

- [1] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. [1](#)
- [2] Fa-Ting Hong and Dan Xu. Implicit identity representation conditioned memory compensation network for talking head video generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23062–23072, 2023. [1](#)
- [3] 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](#)
- [4] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. First order motion model for image animation. *Advances in neural information processing systems*, 32, 2019. [1](#)
- [5] Zhenhui Ye, Tianyun Zhong, Yi Ren, Jiaqi Yang, Weichuang Li, Jiangwei Huang, Ziyue Jiang, Jinzheng He, Rongjie Huang, Jinglin Liu, Chen Zhang, Xiang Yin, Zejun Ma, and Zhou Zhao. Real3d-portrait: One-shot realistic 3d talking portrait synthesis. In *ICLR*, 2024. [1](#)
- [6] 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. [1](#)
- [7] Jian Zhao and Hui Zhang. Thin-plate spline motion model for image animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3657–3666, 2022. [1](#)