

FG-Portrait: 3D Flow Guided Editable Portrait Animation (Supplementary Material)

Yating Xu¹ Yunqi Miao² Evangelos Ververas³ Jiankang Deng³ Jifei Song⁴
¹National University of Singapore ²University of Warwick
³Imperial College London ⁴University of Surrey
{xuyt98, ytswx}@gmail.com {e.ververas16, j.dengl6}@imperial.ac.uk

A. Additional Ablation study

Additional Ablation on Motion Condition. Tab. 1 shows comparison with using driving image as motion condition. Although it shows much better results on the APD and AED, it cheats in the animation by simply copying the driving image as the final output as shown in the Fig. 1. In contrast, we can correctly transfer the motion and maintain the source identity.

Table 1. Comparison of different motion conditions. ‘Dri-Img’ denotes using driving image as the motion condition. ‘S-APD’ and ‘S-AED’ are the APD and AED for self-reenactment task. ‘C-APD’ and ‘C-AED’ are the APD and AED for cross-reenactment task.

Model	S-APD ↓	S-AED ↓	C-APD ↓	C-AED ↓
Dri-Img	1.060	0.131	1.216	0.144
Ours	2.682	0.327	7.764	0.652



Figure 1. Qualitative comparison between ‘Dri-Img’ and Ours. ‘Dri-Img’ directly copies the driving image as the output. In contrast, we can correctly transfer the driving motion to the source person.

Ablation Study on N and δ in the 3D Flow Encoding. Tab. 2 and Tab. 3 show the ablation study of N and δ on the self-reenactment of VFHQ. The performance is generally robust to different combination of N and δ , with a slight degradation when using fewer samples ($N = 10$) or a wider sampling range ($\delta = 0.05m$), due to insufficient sampling density or less accurate 3D flow encoding. Memory of generating 10-frame video mildly increases with larger

N , while remain constant under different δ .

Table 2. Ablation study of N in the self-reenactment on VFHQ.

Method	LPIPS ↓	CSIM ↑	APD ↓	AED ↓	Mem(MB)
N=10	0.164	0.798	2.724	0.332	34110
N=30	0.160	0.807	2.540	0.334	34708
ours	0.158	0.807	2.682	0.327	34402

Table 3. Ablation study of δ in the self-reenactment on VFHQ.

Method	LPIPS ↓	CSIM ↑	APD ↓	AED ↓	Mem(MB)
$\delta = 0.05m$	0.162	0.803	2.742	0.330	34402
$\delta = 0.005m$	0.160	0.804	2.641	0.326	34402
ours	0.158	0.807	2.682	0.327	34402

B. More Qualitative Comparison with Baselines

Fig. 2 shows more qualitative comparisons on testing samples with diverse motion and appearance variations. We show better motion transfer and maintain source identity under these challenging scenarios, *e.g.* identities with long hair, complex accessories, different ethnicities and ages.

C. Animation with Cartoon Portrait

Fig. 3 presents the results on cartoon portraits, where ‘F-Y-E’ denotes Follow-Your-Emoji model. Compared to the baseline, we can more accurately drive the cartoon head. We notice that there is artifact of eyelid closure in Fig. 3 (b). The reason is that our model lacks appropriate appearance priors for cartoon portraits since it is trained exclusively on realistic human portraits. It can be addressed by finetuning on cartoon-specific datasets.

D. Video Results

We provide video results in the supplementary material. Compared to SOTA methods, we can maintain good tem-



Figure 2. Qualitative comparisons on cases with diverse motion and appearance change.

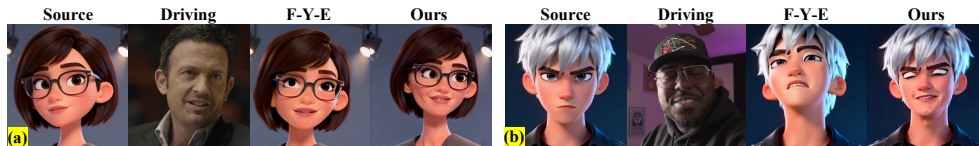


Figure 3. Qualitative comparisons on cartoon portraits. (b) shows a challenging case, which is discussed in Sec. C.

poral consistency and show superior motion transfer at the same time.

E. Temporal consistency Analysis

Tab. 4 shows the temporal consistency analysis on the self-reenactment task of VFHQ. We use Frechet Video Distance (FVD), which is the lower the better temporal consistency. We achieve the second lowest FVD score, which verifies good temporal consistency.

Table 4. Temporal consistency analysis on the self-reenactment task of VFHQ. Best result is marked bold. ‘FYE’ and ‘Hunyuan’ are short for Follow-Your-Emoji and HunyuanPortrait, respectively.

Method	EMOPortrait	X-Portrait	FYE	Face-Adapter	Hunyuan	Ours
FVD	567.2	575.3	382.6	472.3	430.2	412.1