MVPortrait: Text-Guided Motion and Emotion Control for Multi-view Vivid Portrait Animation

Supplementary Material

1. Training

The loss for FLAME2Video stage is formulated as

$$\mathbb{E}_{\epsilon\sim\mathcal{N}(\mathbf{0},\mathbf{I}),\mathbf{x}_t,\mathbf{c},t}[\|\epsilon-\epsilon_\theta(\mathbf{x}_t;\mathbf{c},t)\|_2^2] \quad (1)$$

where $\mathbf{c}$ is image embedding encoded by CLIP and $\mathbf{x}_t$ is derived by adding t-step noises to $\mathbf{x}_0$, and ϵ and ϵ_θ respectively represent the ground truth and predicted noise by Denoising UNet.

In the Text2FLAME stage, the conditional learning strategy we use is a classifier-free guidance. We set the text condition to null with a probability of 10%. Following [7], we set a maximum frame limit during the training, with padding applied to shorter videos. The diffusion sampling step is set to 1000. In the FLAME2Video stage, the diffusion sampling step is set to 25. We utilize the pre-trained UNet weights from runwayml/stable-diffusion-v1-5.

2. Metrics

In this section, we explain the calculation processes for the metrics not detailed in the main text.

Variability. We devise a metric to evaluate the amplitude of head movements to assess the expressiveness of portrait videos. Specifically, DECA [4] is employed to estimate FLAME parameters, through which the yaw, pitch, and roll angles are extracted from the head pose. The mean variability of motion in the generated videos is quantified by calculating the temporal differences between consecutive frames for these pose angles.

FLAME-L1. To quantify the differences in pose and expression between the generated and driving videos during cross-reenactment, we utilize DECA [4] to estimate FLAME parameters, including pose and expression, for both the generated and driving videos. The FLAME-L1 metric is then computed by calculating the L1 difference between the corresponding parameters.

3. Quantitative Comparison

3.1. Comparison with more baselines

We compare with text-guided video generative models like CogVideoX[9] and DynamiCrafter[8] to benchmark our model's performance against current leading methods. We fine-tune them on CelebV-Text, with test results in Tab. 1.

3.2. Out-of-Distribution Performance

We generate 10 out-of-distribution (OOD) text descriptions using GPT-4, and present the quantitative results for these cases in Tab. 1, where a performance drop is observed. However, the example in Fig. 1 illustrates that our model shows some generalization capability.

3.3. Multi-view Comparison

For fairness, we fine-tune Triplanenet [1] and Portrait4D-v2 [3] on RenderMe-360 [6] training set, and present the evaluation results on the RenderMe-360 test set in Tab. 2. Multi-view accuracy is measured as the average LPIPS, SSIM, and L1 differences between the predictions and ground truth across all viewpoints.

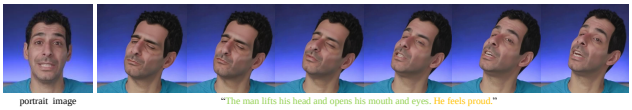

Figure 1. The example for demonstrating out-of-distribution performance.

4. Qualitative Comparison

We present additional visual comparisons to provide readers with a clearer view of the differences between various methods. Since FLAME [5] acts as an intermediate representation, our framework becomes the first to support text, video, and audio-driven portrait animations. In the following, we present qualitative comparisons of our method with others under different driving signals.

4.1. Text-driven Animation

For text-driven portrait animation, we compare our method with AnimateAnything [2] and MMVID-interp [11], both of which were fine-tuned on the CelebV-Text dataset to ensure a fair comparison. As end-to-end generation models, AnimateAnything and MMVID-interp achieve text-driven portrait video generation by learning an implicit mapping between text and video. However, both methods exhibit weaker controllability compared to our approach, which leverages FLAME for explicit control. This advantage is primarily due to our use of a text-guided diffusion model to generate the corresponding head poses and facial expressions. As shown in Fig. 2, our method surpasses the other two methods in terms of motion and emotion consistency with the text description, and demonstrates superior vividness. Furthermore, videos are provided below for readers to review.

Method	LIQE↑	FID↓	FVD↓	CLIPSIM↑	VideoClip↑	Variability↑	MC↑	EC↑	VS↑
DynamiCrafter-ft	4.306	80.9	**552.8**	0.172	0.557	0.074	1.70	1.22	1.78
CogVideoX-ft	4.046	171.7	962.3	0.179	0.589	0.119	2.37	2.22	2.25
MVPortrait	**4.760**	**28.6**	570.0	**0.183**	**0.595**	0.110	**2.57**	**2.29**	**2.48**
No smoothing	3.849	108.4	1213.5	0.169	0.586	**0.243**	1.88	1.22	2.13
Window size: 5	4.590	67.0	948.7	0.180	0.593	0.104	2.25	2.10	2.11
Joint training	4.409	76.8	824.7	0.175	0.559	0.068	1.50	1.44	1.56
OOD cases	4.245	-	-	0.175	0.561	0.111	1.67	2.11	1.89

Table 1. Comparison of text-guided animation on the CelebV-Text test set.

Method	LPIPS↓	SSIM↑	L1↓	ID↑
Triplanenet-ft	0.3752	0.5111	0.2200	0.8803
Portrait4D-v2-ft	0.3725	0.5150	0.1957	0.8342
MVPortrait (view number: 4)	**0.2445**	**0.5512**	**0.1735**	**0.8891**
view number: 2	0.3206	0.5370	0.1971	0.8224
w/o view attention (view number: 1)	0.3959	0.4624	0.2746	0.7826

Table 2. Comparison of multi-view portrait synthesis on the RenderMe-360 test set.

4.2. Video-driven Animation

Given a driving video, we first use the FLAME estimation method, DECA [4] in our experiment to estimate the corresponding FLAME sequence and obtain the renderings. Next, we use these renderings to generate the animated video. Our experiments reveal that videos generated by FollowYourEmoji exhibit significant flickering artifacts, as evident in the video. While LivePortrait demonstrates strong driving performance, the pose and expression in its generated videos often fail to align with those in the driving video. In contrast, our method produces videos with superior robustness and controllability.

Visual comparisons are provided in Fig. 3. In each subplot, the first row shows the driving video, and the second row shows the FLAME sequence, which is constructed from the FLAME pose and expression parameter sequences estimated from the driving video, along with the facial shape parameters and facial detail vectors estimated from the reference image. Thus, the FLAME sequence here represents both the head movements and facial expressions exhibited in the driving video, as well as the facial shape and details of the reference portrait. These FLAME sequences can serve as a reference for evaluating the effectiveness of the driving algorithm. It is clear from subplots (a) and (b) that the head pose in the results generated by LivePortrait is significantly inconsistent with the driving video. Additionally, in subplot (c), severe blurring is observed in the video generated by LivePortrait, which may be due to the presence of hands in the driving video, a scenario that LivePortrait struggles to handle robustly. FollowYourEmoji also struggles to handle head pose and tends to produce larger mouth movements compared to the driving video, as shown in all subplots. We encourage readers to watch the videos in the supplementary materials to gain a more intuitive understanding of the differences between the methods.

4.3. Audio-driven Animation

In our framework, audio-driven generation is also carried out in two stages. In the first stage, we use an audio-driven head generation method, TalkShow [10] in our experiment, to produce FLAME parameters. The FLAME parameters are used for generating FLAME images as guidance conditions. In the second stage, FLAME is used to create the animation. When showcasing MVPortrait's performance in audio-driven animation, we also present the FLAME generated by TalkShow, to better assess the synchronization between audio and video. Refer to Fig. 4 for comparison.

5. Ablation

5.1. Text2FLAME

We ablate the Text2FLAME stage using three variants: *No Smoothing*, and *Larger Network Size*, *Joint Generation*. As mentioned in the main text, the No Smoothing variant causes mismatched expressions and head jitter, the Larger Network Size variant generates correct expressions but lacks head movement, and the Joint Generation variant shows incorrect expressions and static motion. We include

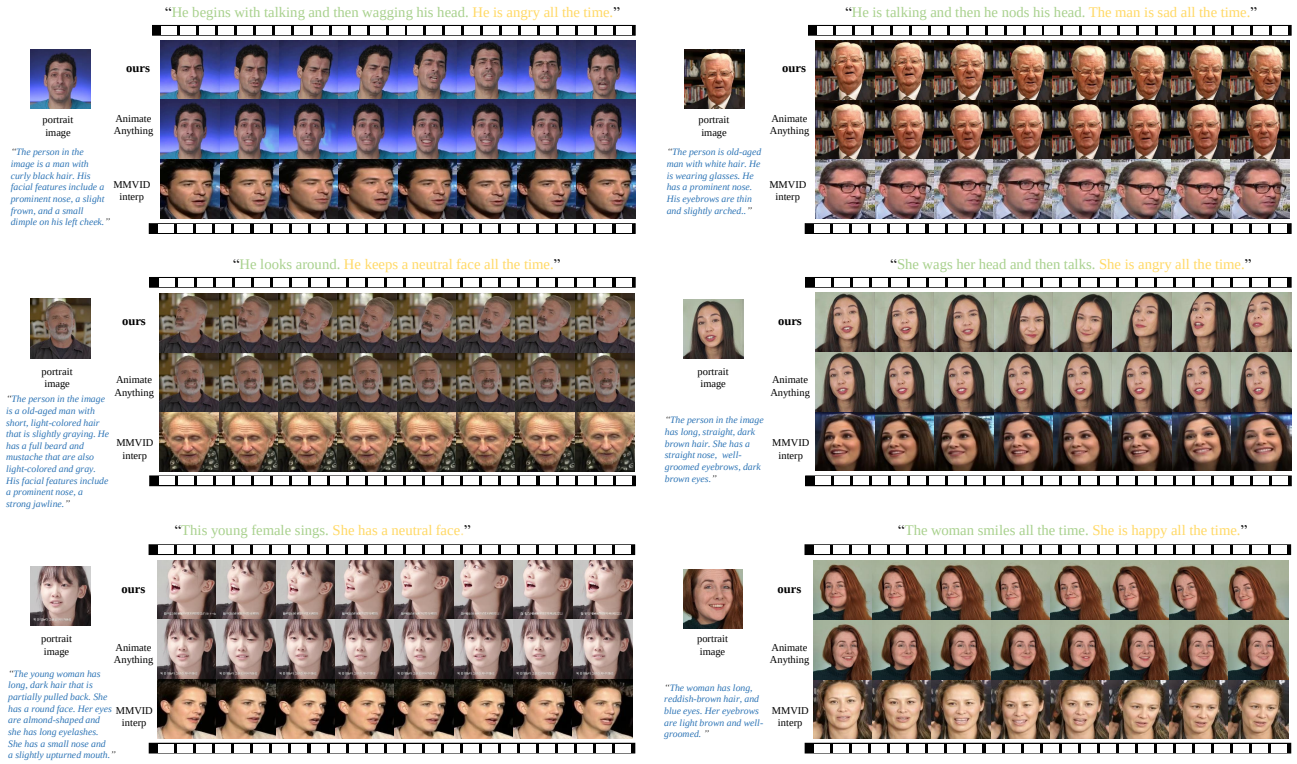

Figure 2. The qualitative comparison of text-guided portrait animation. The motion descriptions are highlighted in green, while the emotion descriptions are marked in yellow. The blue text represents the appearance descriptions, which is only used by MMVID-interp. The generated video frames are displayed sequentially from left to right.

animations of these variants in the supplementary materials. Please watch them for comparison.

Additionally, for smoothing operations, we conduct an additional experiment with a sliding window size of 5. The quantitative results are shown in Tab. 1. Note that the window size for MVPortrait is 3, while the window size for *no smoothing* is 1. This ablation shows a window size of 3 balances stability and naturalness best.

5.2. FLAME2Video

In this part, we present additional multiview results to demonstrate the effectiveness of our view attention mechanism in Fig. 5. We train a 2-view model and report evaluation results in Tab. 2, which shows that the performance improves as the number of views increases, up to the maximum of 4 supported on an 80GB A100 GPU.

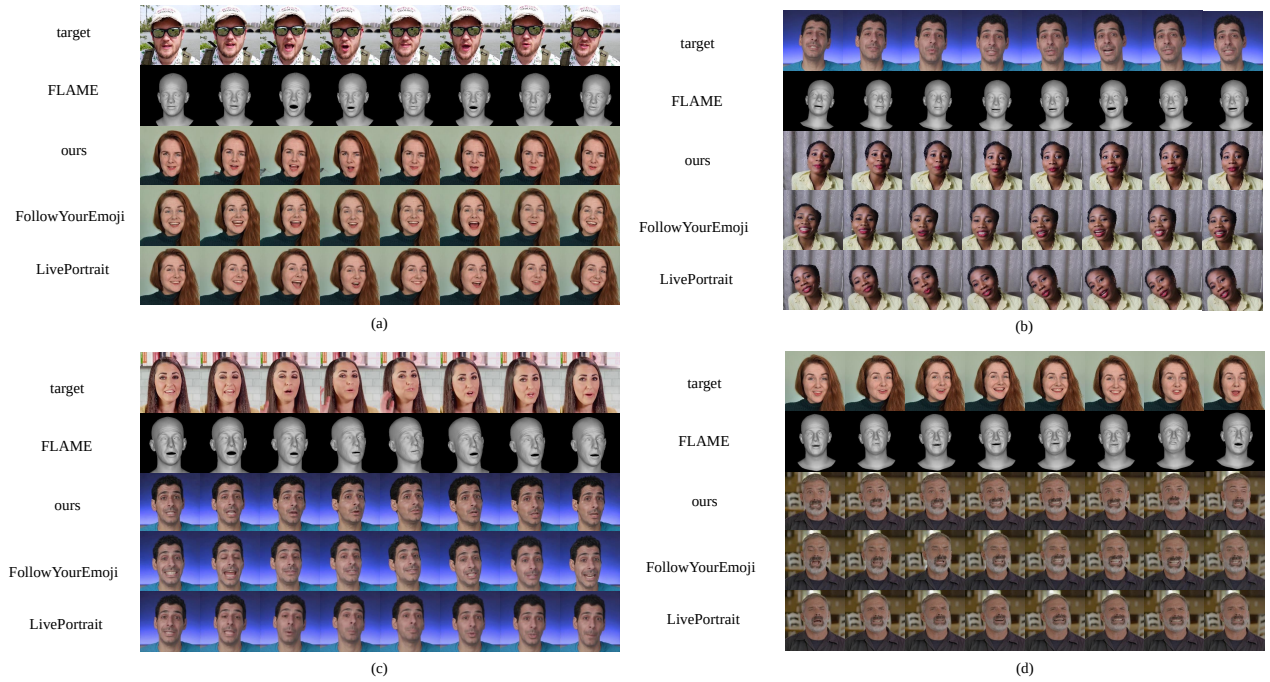

Figure 3. The qualitative comparison of video-driven portrait animation. The generated video frames are displayed sequentially from left to right.

Figure 4. The qualitative comparison of audio-driven portrait animation. The generated video frames are displayed sequentially from left to right.

Figure 5. The qualitative ablation of view attention.

References

[1] Ananta R Bhattarai, Matthias Nießner, and Artem Sevastopolsky. Triplanenet: An encoder for eg3d inversion. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3055–3065, 2024. 1

[2] Zuozhuo Dai, Zhenghao Zhang, Yao Yao, Bingxue Qiu, Siyu Zhu, Long Qin, and Weizhi Wang. Animateanything: Fine-grained open domain image animation with motion guidance, 2023. 1

[3] Yu Deng, Duomin Wang, and Baoyuan Wang. Portrait4d-v2: Pseudo multi-view data creates better 4d head synthesizer. *arXiv preprint arXiv:2403.13570*, 2024. 1

[4] Yao Feng, Haiwen Feng, Michael J Black, and Timo Bolkart. Learning an animatable detailed 3d face model from in-the-wild images. *ACM Transactions on Graphics (ToG)*, 40(4): 1–13, 2021. 1, 2

[5] Tianye Li, Timo Bolkart, Michael. J. Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4D scans. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36(6):194:1–194:17, 2017. 1

[6] Dongwei Pan, Long Zhuo, Jingtan Piao, Huiwen Luo, Wei Cheng, Yuxin Wang, Siming Fan, Shengqi Liu, Lei Yang, Bo Dai, Ziwei Liu, Chen Change Loy, Chen Qian, Wayne Wu, Dahua Lin, and Kwan-Yee Lin. Renderme-360: A large digital asset library and benchmarks towards high-fidelity head avatars. *Advances in Neural Information Processing Systems*, 36, 2024. 1

[7] Guy Tevet, Sigal Raab, Brian Gordon, Yoni Shafir, Daniel Cohen-or, and Amit Haim Bermano. Human motion diffusion model. In *The Eleventh International Conference on Learning Representations*, 2023. 1

[8] Jinbo Xing, Menghan Xia, Yong Zhang, Haoxin Chen, Wangbo Yu, Hanyuan Liu, Gongye Liu, Xintao Wang, Ying Shan, and Tien-Tsin Wong. Dynamicrafter: Animating open-domain images with video diffusion priors. In *European Conference on Computer Vision*, pages 399–417. Springer, 2024. 1

[9] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024. 1

[10] Hongwei Yi, Hualin Liang, Yifei Liu, Qiong Cao, Yandong Wen, Timo Bolkart, Dacheng Tao, and Michael J Black. Generating holistic 3d human motion from speech. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 469–480, 2023. 2

[11] Jianhui Yu, Hao Zhu, Liming Jiang, Chen Change Loy, Weidong Cai, and Wayne Wu. CelebV-Text: A large-scale facial text-video dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 1