Supplementary Material for FlashPortrait

Shuyuan Tu[1] Yueming Pan[3] Yinming Huang[1] Xintong Han[4] Zhen Xing[5]
Qi Dai[2] Kai Qiu[2] Chong Luo[2] Zuxuan Wu[1]
[1]Fudan University [2]Microsoft Research Asia [3]Xi'an Jiaotong University
[4]Tencent Inc. [5]Tongyi Lab, Alibaba Group
https://francis-rings.github.io/FlashPortrait

1. Evaluation Metrics

Following previous portrait animation evaluation settings, we implement numerous quantitative evaluation metrics, including FID, FVD, LMD, AED, APD, and MAE, to compare our FlashPortrait with current state-of-the-art portrait animation models. The details of the above metrics are described as follows:

(1) FID refers to measure the similarity in feature distribution between synthesized and real images, employing Inception v3 features.
(2) FVD refers to evaluate temporal coherence through features extracted from a pretrained model [9].
(3) LMD refers to measure the accuracy of synthesized facial expressions. The landmarks are extracted using Mediapipe. It computes the average Euclidean distance between the facial landmarks of the reference and synthesized images.
(4) AED refers to calculate the Manhattan distance of expression from SMIRK [8], with lower values indicating better expression.
(5) APD calculate the Manhattan distance of pose parameters from SMIRK [8], with lower values indicating better pose similarity.
(6) MAE refers to measure the Mean Angular Error on the eye movement accuracy.

2. Preliminaries

Diffusion models function through a stochastic process, consisting of two main phases: a forward diffusion step and a reverse denoising step for controlled noise addition and removal. In the forward process, Gaussian noise is gradually introduced to a data sample $\boldsymbol{x}_0 \sim \boldsymbol{p}_{\text{data}}$, where $\boldsymbol{p}_{\text{data}}$ represents the underlying data distribution. This is done as follows, based on the Rectified Flow method [7]:

$$\boldsymbol{x}_t = (1-t)\boldsymbol{x}_0 + t\boldsymbol{x}_1, \tag{1}$$

where $t \in [0, 1]$ denotes the timestep. After $\boldsymbol{T}$ diffusion steps, the original data sample $\boldsymbol{x}_0$ is transformed into pure Gaussian noise $\boldsymbol{x}_1 \sim \mathcal{N}(0, I)$. In the reverse denoising process, the diffusion model $\boldsymbol{\varepsilon}_\theta(\boldsymbol{x}_t, t)$ is trained to predict the velocity $(\boldsymbol{x}_1 - \boldsymbol{x}_0)$ conditioned on the noisy latents x_t and the timestep t. To train the model, the Mean Squared Error (MSE) loss is applied:

$$\mathcal{L} = \mathbb{E}_{\boldsymbol{x}_0, \boldsymbol{\varepsilon}, t}(\|\boldsymbol{\varepsilon}_\theta(\boldsymbol{x}_t, t) - (\boldsymbol{x}_1 - \boldsymbol{x}_0)\|^2). \tag{2}$$

This framework ensures accurate denoising, gradually recovering the original data from noisy latents.

Table 1. User preference of FlashPortrait compared with other competitors. Higher indicates users prefer more to our model.

Model	L-A	A-A	B-A	I-A
LivePortrait [5]	95.4%	97.2%	98.5%	97.9%
HunyuanPortrait [13]	94.8%	96.4%	98.2%	97.6%
FantasyPortrait [10]	95.2%	95.8%	97.7%	96.8%
Wan-Animate [1]	92.8%	93.7%	97.4%	96.5%

3. Discussion on AdaIN

AdaIN [6] and our FlashPortrait share a similar formulation but different goals. AdaIN performs global feature replacement, whereas NFEB serves as a cross-modal manifold calibrator that bridges the distribution gap between facial embeddings and latents.

4. User Study

To assess perceptual quality in a subjective way, we conducted a user study involving 30 curated samples. The participants, primarily university students and faculty, are first shown the reference image and its driven video. They then view two synthesized results—one from FlashPortrait and another from a competing method—presented in random order. Participants are then asked to answer the following questions: L-A/A-A/B-A/I-A: "Which one has better facial expression motion/foreground appearance/background/identity alignment with the driven video/reference". The results in Table 1 demonstrate that FlashPortrait is consistently preferred in all aspects.

Table 2. GPU memory consumption comparison.

Mode	Mem	Mode	Mem
TaylorSeer	18G	Ours(K=2,n=3)	18G
Self-Forcing	10G	Ours(K=8,n=3)	18G
Ours(K=5,n=3)	18G	Ours(K=5,n=1)	12G

5. GPU Memory Consumption

The baseline cost is 10G on 4 A100. The GPU memory consumption comparison results are shown in Table 2. Distillation-based self-forcing incurs no extra GPU cost, while our FlashPortrait and TaylorSeer require caching and thus increase cost. As K only affects the prediction horizon, it does not change the number of cached latents and thus does not add cost. n controls the number of caches, increasing cost.

6. Implementation and Dataset Details

We train the model using AdamW ($\beta_1 = 0.9$, $\beta_2 = 0.999$) and run the entire optimization in bfloat16. Distributed data parallelism is handled through DeepSpeed-Stage-3, which manages gradient synchronization and memory efficiency during training.

In terms of the training dataset, our training dataset consists of three parts, including Hallo3 [3], Celebv-HQ [14], and collected videos from the internet (BilBil, YouTube, and TikTok). We utilize the Q-Align [11] to filter for higher-quality videos by assessing the overall video fidelity. We also apply InsightFace [4] to filter out videos with a facial confidence score below 0.8. We obtain the final training dataset, containing roughly 2000 hours of videos.

Regarding the testing dataset, we first randomly select 100 videos (5-20 seconds long) from Voxceleb2 [2] and Vfhq [12] to construct the first simple testing dataset. To validate the robustness of our FlashPortrait, we further select 100 unseen videos (1-3 minutes long, FPS=30) from the internet to construct the testing dataset Hard100. Some examples are shown in Fig. 1. The sources of Hard100 come from various social media platforms, such as BilBil, YouTube, and TikTok. The selected videos span both indoor and outdoor environments, and the protagonists exhibit substantial demographic diversity, including balanced distributions across gender and ethnicity. The videos contain both upper-body and full-body subjects, with actions ranging from simple standing poses to complex interactions with objects in the scene. Consequently, our curated testing dataset is substantially more challenging than existing open-source testing datasets (Voxceleb2 [2] and Vfhq [12]) in terms of subject diversity, environmental diversity, and pose variability. Moreover, the average duration of our selected videos is approximately two minutes, which is significantly longer than that of existing open-source testing datasets.

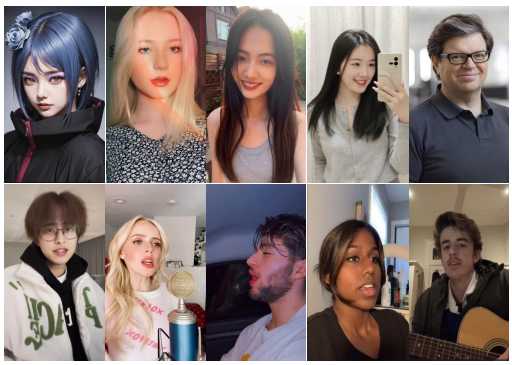

Figure 1. Examples from Hard100.

Table 3. Ablation study on different weight assignment.

Model	AED↓	APD↓	MAE↓	Speed↓
w/o $\boldsymbol{s}(t)$	34.73	28.44	16.12	709s
w/o $\boldsymbol{w}(t,l,i)$	40.48	33.52	18.75	**688s**
Ours	**29.68**	**24.40**	**12.54**	720s

7. Additional Ablation on Acceleration

We conduct an ablation study on two dynamic functions in our proposed Adaptive Latent Prediction Acceleration Mechanism, as shown in Table 3. We observe that removing $\boldsymbol{s}(t)$ and $\boldsymbol{w}(t,l,i)$ significantly degrades performance. It indicates that $\boldsymbol{s}(t)$ and $\boldsymbol{w}(t,l,i)$ can facilitate the accuracy of predicted latents based on the latent variation rate at particular timesteps and the derivative magnitude ratio among diffusion layers. The underlying reason is that $\boldsymbol{s}(t)$ and $\boldsymbol{w}(t,l,i)$ jointly regulate the approximation between $\triangle^i \boldsymbol{f}(t,l)$ and $\boldsymbol{f}^{(i)}(t,l)$, ensuring robustness of latent prediction across diverse scenarios, even when the generated videos exhibit large motion variations.

We further conduct an ablation study on different acceleration methods, presenting the results through progressive visualizations, as shown in Fig.2. We observe that as the number of generated frames increases, all competitors become progressively unstable, particularly in terms of facial and background consistency. When the sequence length exceeds 800 frames, all competitors exhibit varying degrees of face and body distortion, as well as color drift. Moreover, the generated portrait no longer strictly follows the driven video, with facial expressions turning stochastic, especially in mouth closure, eye motion, and head rotation. By contrast, our FlashPortrait achieves a $6\times$ inference speedup over the baseline while maintaining comparable visual quality and preserving high-fidelity identity consistency. Moreover, the generated facial expressions strictly follow the guidance of the driven video, which demonstrates the superiority of our Adaptive Latent Prediction Acceleration Mechanism over previous acceleration methods in the long-length portrait animation.

8. Full/Half Body Portrait Animation

We perform a qualitative experiment in full/half-body portrait animations, as shown in Fig. 3. Each reference image has a complex background layout and intricate foreground appearance. The first case even involves interactions with objects from the environment, such as an instrument, making it more challenging to maintain identity consistency and facial expression synchronization with the driven video. We can see that our FlashPortrait has the capacity to synthesize full/half-body portrait animations, even involving interactions with external objects.

9. Long Portrait Animation

To further validate the performance of our FlashPortrait in long-length portrait animation, we perform a qualitative experiment in an extremely long case (4 minutes, FPS=30), as shown in Fig.4. Our FlashPortrait can still maintain identity consistency and ensure expression synchronization with the driven video, even after synthesizing 7000+ frames. From a theoretical perspective, FlashPortrait can synthesize infinite-length high-quality identity-preserving animations.

10. More Portrait Animation

Fig. 5, Fig. 6, Fig. 7, Fig. 8, and Fig. 9 presents additional portrait animation result synthesized by our FlashPortrait. Each driven video contains 1800+ frames, and we only select synthesized frames from the last 100 frames for presentation. The reference protagonists exhibit rich diversity, encompassing both male and female subjects across various ethnicities. They also present complex visual characteristics, including intricate hairstyles, richly textured clothing, elaborate tattoo patterns, and a wide range of refined accessories. Each driven video contains substantial and dynamic facial expression motions with irregular expression patterns, such as head rotations and rapid blinking. We can observe that our FlashPortrait can accurately animate the reference image based on the driven video while maintaining strong identity consistency even after synthesizing 1800 frames. For example, the third row of Fig. 7 contains dramatic facial expression motions and exaggerated expression patterns, making it challenging for model to preserve ID. Our FlashPortrait can still accurately manipulate lip movement, eye movement, and head movement of the reference while maintaining high-quality ID consistency.

11. Limitation and Future Work

Fig. 10 shows one failure case of our FlashPortrait. When the reference protagonist is a humanoid character, such as a game avatar or a mythological figure, its appearance does not strictly conform to real human facial standards. Since our model is primarily trained on real human video data, FlashPortrait tends to synthesize a more realistic human face to replace the original reference protagonist's face. This adaptation disrupts identity consistency and results in generated faces that deviate substantially from the reference image. One potential solution is to introduce an additional reference network to explicitly capture the face details of the reference images. This reference network needs to be trained from scratch on large-scale diverse video datasets. This part is left as future work.

References

[1] Gang Cheng, Xin Gao, Li Hu, Siqi Hu, Mingyang Huang, Chaonan Ji, Ju Li, Dechao Meng, Jinwei Qi, Penchong Qiao, et al. Wan-animate: Unified character animation and replacement with holistic replication. *arXiv preprint arXiv:2509.14055*, 2025. 1

[2] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. Voxceleb2: Deep speaker recognition. *arXiv preprint arXiv:1806.05622*, 2018. 2

[3] Jiahao Cui, Hui Li, Yun Zhan, Hanlin Shang, Kaihui Cheng, Yuqi Ma, Shan Mu, Hang Zhou, Jingdong Wang, and Siyu Zhu. Hallo3: Highly dynamic and realistic portrait image animation with video diffusion transformer. In *CVPR*, 2025. 2

[4] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *CVPR*, 2019. 2

[5] Jianzhu Guo, Dingyun Zhang, Xiaoqiang Liu, Zhizhou Zhong, Yuan Zhang, Pengfei Wan, and Di Zhang. Liveportrait: Efficient portrait animation with stitching and retargeting control. *arXiv preprint arXiv:2407.03168*, 2024. 1

[6] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *ICCV*, 2017. 1

[7] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022. 1

[8] George Retsinas, Panagiotis P Filntisis, Radek Danecek, Victoria F Abrevaya, Anastasios Roussos, Timo Bolkart, and Petros Maragos. 3d facial expressions through analysis-by-neural-synthesis. In *CVPR*, 2024. 1

[9] Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphaël Marinier, Marcin Michalski, and Sylvain Gelly. Fvd: A new metric for video generation. 2019. 1

[10] Qiang Wang, Mengchao Wang, Fan Jiang, Yaqi Fan, Yonggang Qi, and Mu Xu. Fantasyportrait: Enhancing multi-character portrait animation with expression-augmented diffusion transformers. *arXiv preprint arXiv:2507.12956*, 2025. 1

[11] Haoning Wu, Zicheng Zhang, Weixia Zhang, Chaofeng Chen, Liang Liao, Chunyi Li, Yixuan Gao, Annan Wang, Erli Zhang, Wenxiu Sun, et al. Q-align: Teaching lmms for visual scoring via discrete text-defined levels. *arXiv preprint arXiv:2312.17090*, 2023. 2

[12] Liangbin Xie, Xintao Wang, Honglun Zhang, Chao Dong, and Ying Shan. Vfhq: A high-quality dataset and benchmark for video face super-resolution. In *CVPR*, 2022. 2

Figure 2. Ablation study on different acceleration methods. *w/o* DF refers to *w/o* Dynamic Functions.

[13] Zunnan Xu, Zhentao Yu, Zixiang Zhou, Jun Zhou, Xiaoyu Jin, Fa-Ting Hong, Xiaozhong Ji, Junwei Zhu, Chengfei Cai, Shiyu Tang, et al. Hunyuanportrait: Implicit condition control for enhanced portrait animation. In *CVPR*, 2025. 1

[14] Hao Zhu, Wayne Wu, Wentao Zhu, Liming Jiang, Siwei Tang, Li Zhang, Ziwei Liu, and Chen Change Loy. CelebV-HQ: A large-scale video facial attributes dataset. In *ECCV*, 2022. 2

Figure 3. Full/Half-body portrait animation results. The images with red borders are the reference images.

Figure 4. Long portrait animation results. The images with red borders are the reference images.

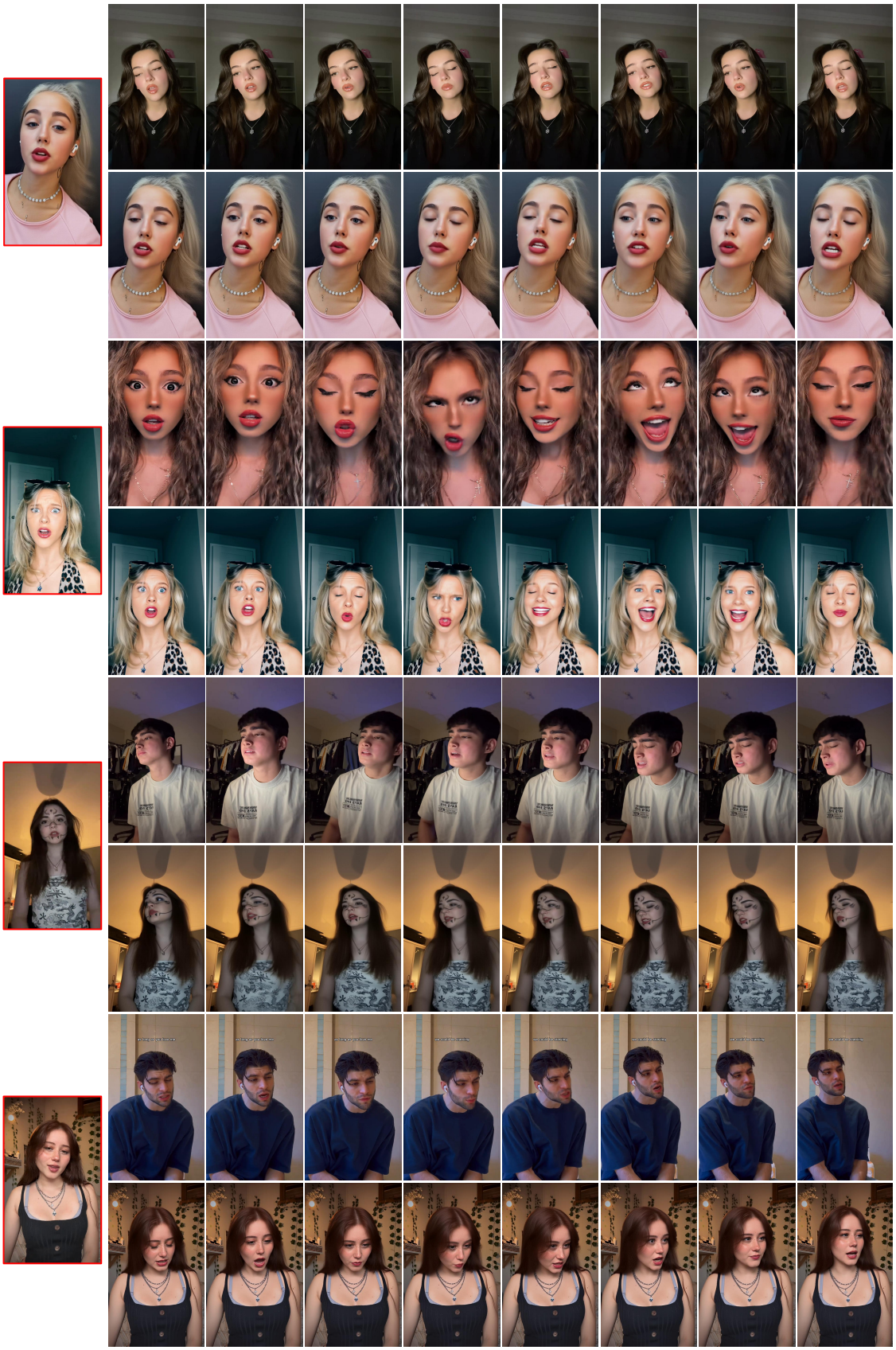

Figure 5. portrait animation results (1/5). The images with red borders are the reference images.

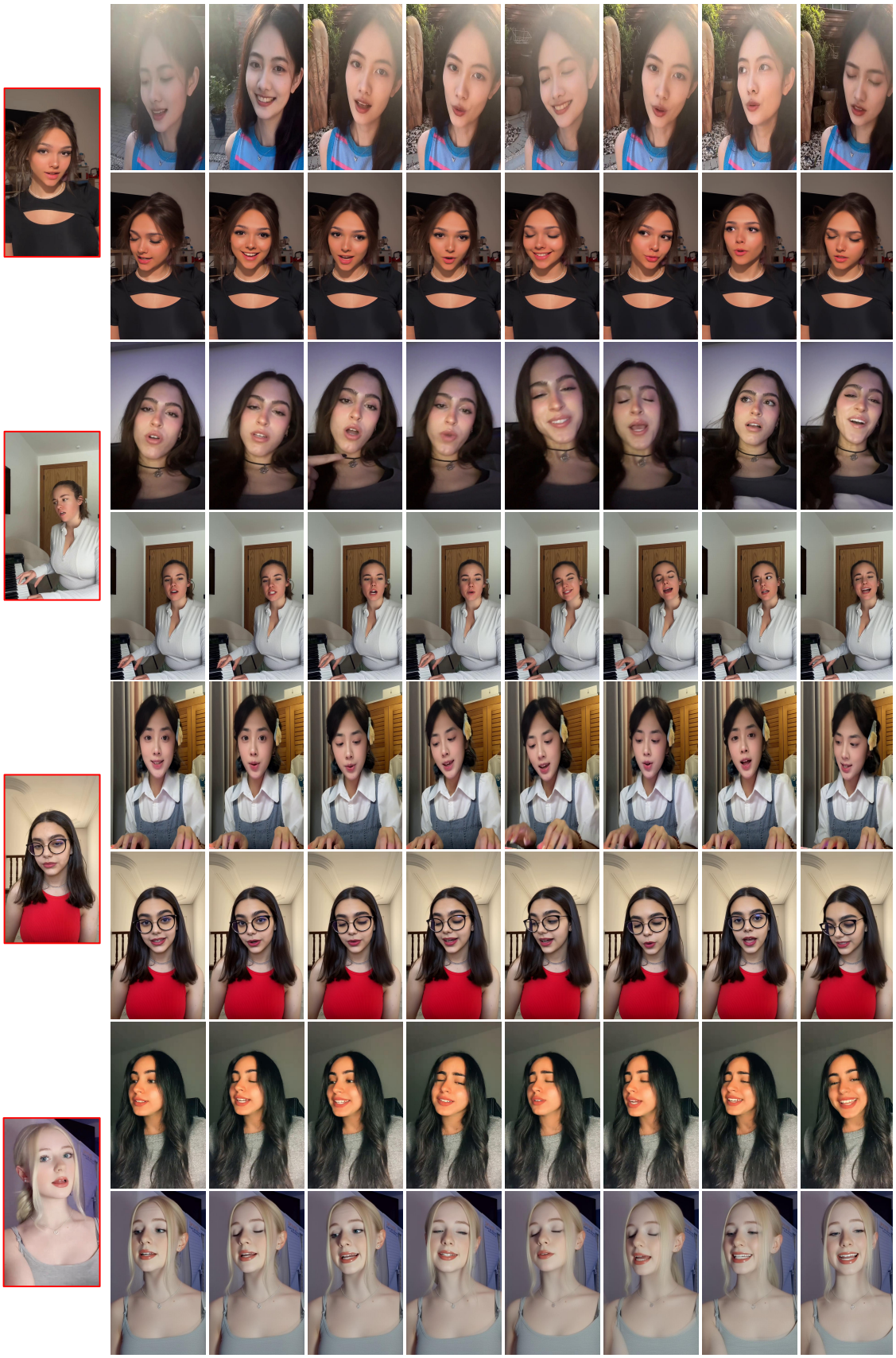

Figure 6. portrait animation results (2/5). The images with red borders are the reference images.

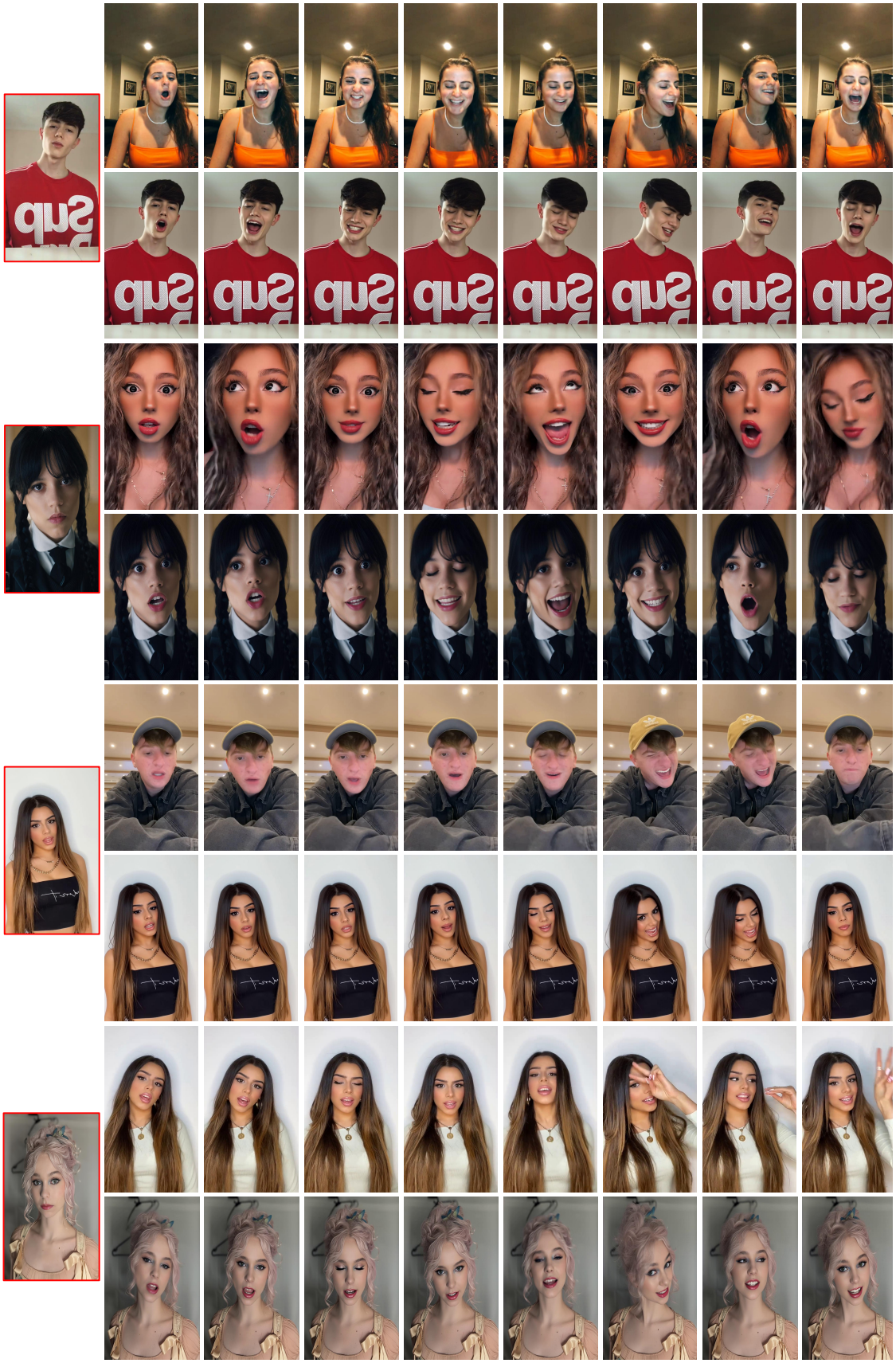

Figure 7. portrait animation results (3/5). The images with red borders are the reference images.

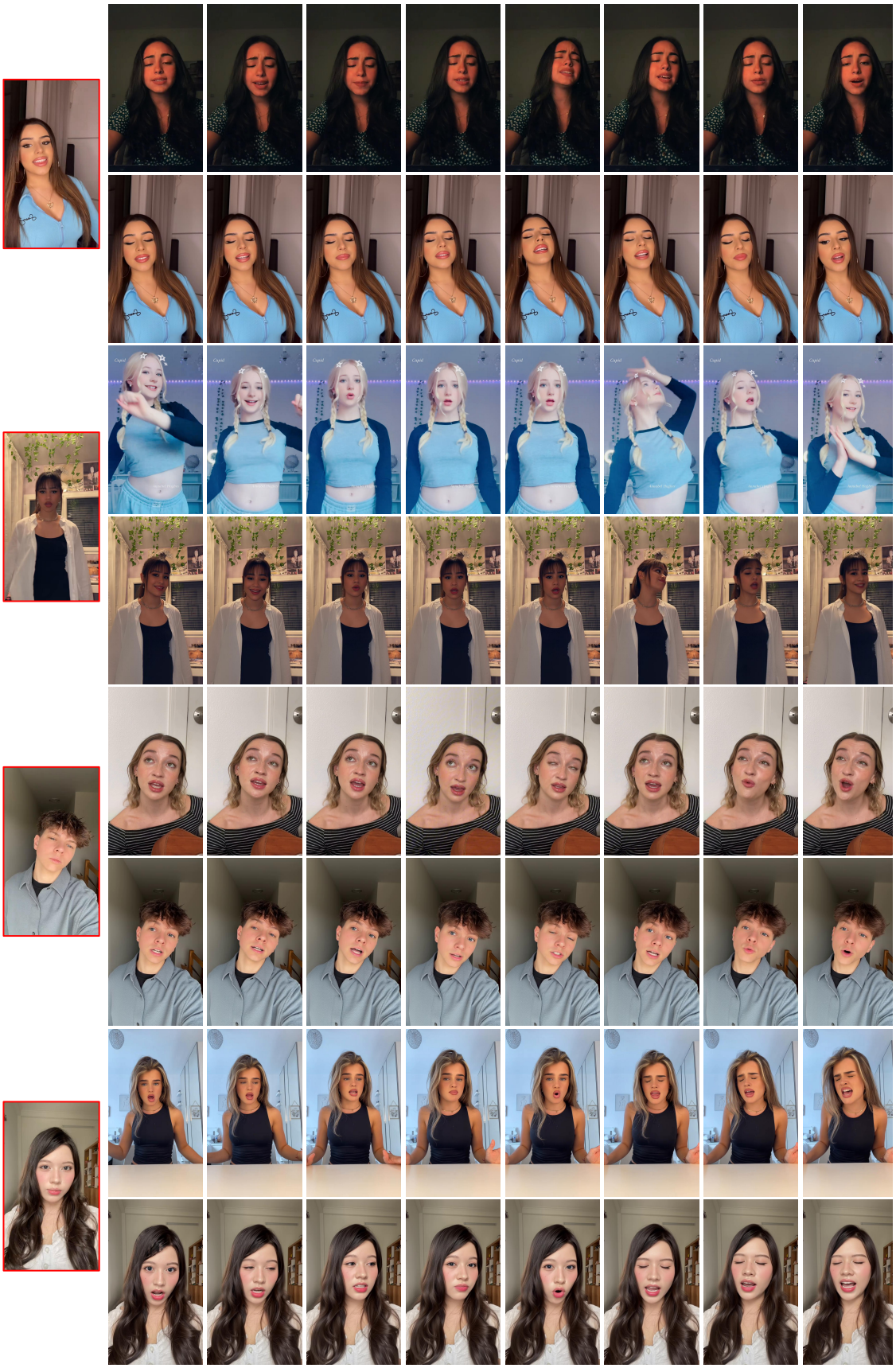

Figure 8. portrait animation results (4/5). The images with red borders are the reference images.

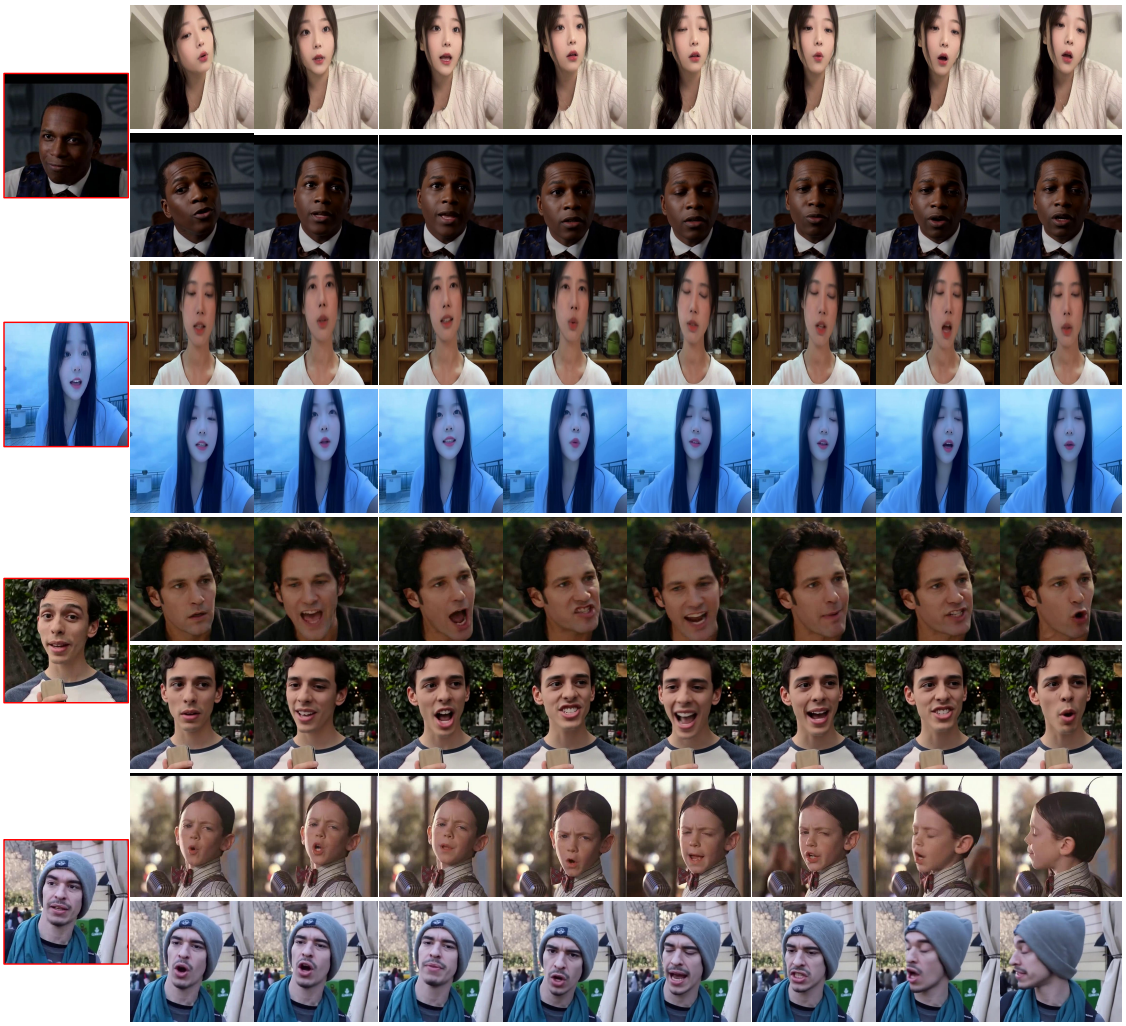

Figure 9. portrait animation results (5/5). The images with red borders are the reference images.

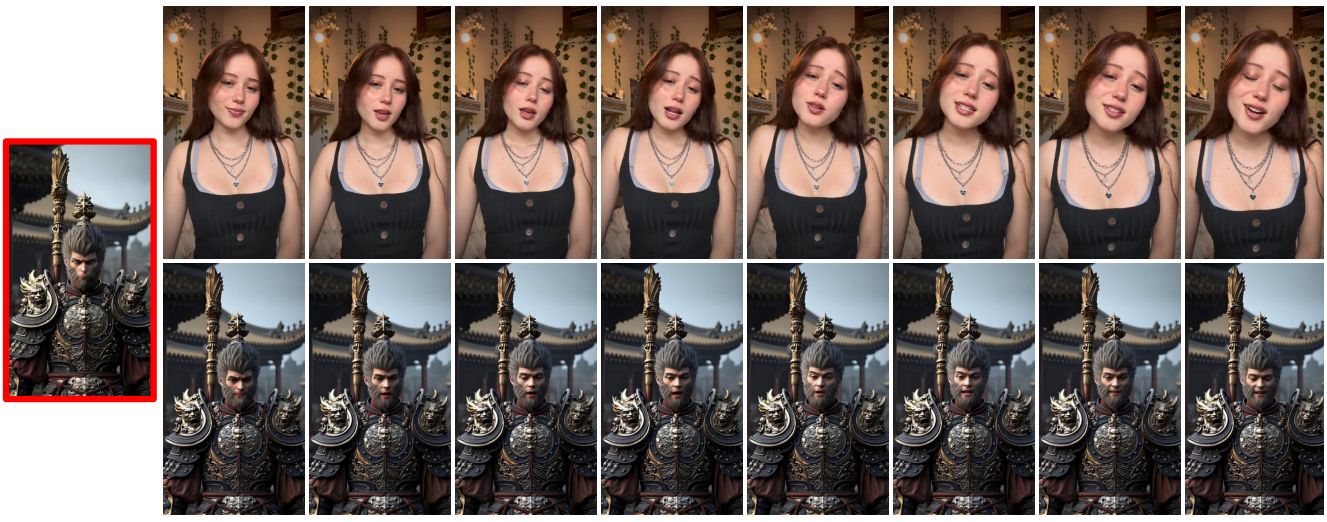

Figure 10. One failure case of our FlashPortrait. The images with red borders are the reference images.