EFHQ: Multi-purpose ExtremePose-Face-HQ dataset

Supplementary Material

9. More on dataset pipeline

9.1. Pose bin

Fig. 11 illustrates the distribution of frames per bin following our automated processing pipelines, presented on a logarithmic scale. As depicted in the figures, the original datasets contain a substantial volume of frames, yet a predominant portion consists of frontal faces. This observation substantiates our previously articulated arguments regarding the prevalent bias toward frontal views. Leveraging our preprocessing pipelines enables us to concentrate solely on extreme view images, reserving the utilization of frontal views for some specific subtasks like face reenactment and face verification. Furthermore, Fig. 12 shows the pose distribution in terms of angle bins of the original two video-dataset compared to IJB-C and CPLFW.

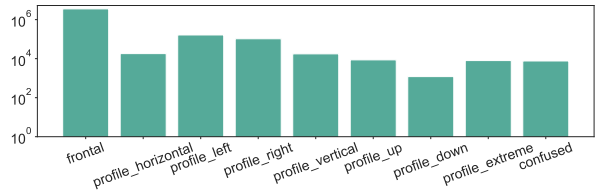

Figure 11. **Frame distribution** for each pose bin after our automated attributes preparation process.

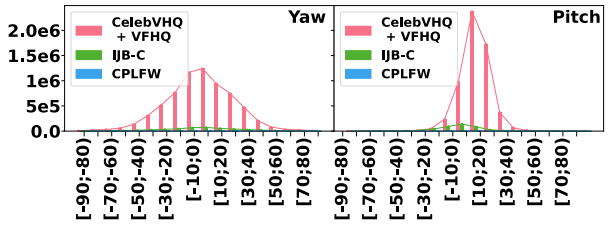

Figure 12. **Pose distribution** in terms of angle bins of CelebV-HQ + VFHQ, IJB-C, CPLFW

9.2. Model hyperparameters

In the process of extracting face bounding boxes and 5 keypoints via RetinaFace[36], we keep only detected faces with confidence above 0.9 to ensure the retention of profile faces while minimizing false negatives and low quality facial detections. Additionally, faces with an area smaller than 256×256 in both datasets are disregarded. To further refine the selection, the HyperIQA score, specifically set at 42 according to [48], is employed during the sampling phase to exclude frames of poor quality. Moreover, to address the issue of low-light images, which is not explicitly assessed by HyperIQA, a method referenced in [3] is utilized, filtering out images with a brightness value below 40, a value determined empirically to suit our specific requirements. For DirectMHP, we first apply multiscale inference and later use a confidence threshold of 0.7 and an IoU threshold of 0.45 during the non max suppression (NMS) algorithm. All other models, if not specified, follow the original setup of the original authors.

10. Dicussion on Negative Societal Impacts

Our dataset comes from putting together two public datasets [48, 54]. Consequently, we share similar concerns regarding privacy and bias with the original datasets from which our data is sourced. It is crucial for any user of our dataset to adhere strictly to the rules outlined in the original datasets regarding its usage. Moreover, we advocate and encourage further exploration and endeavor aimed at developing strategies that proactively minimize and counteract any deleterious societal implications, thereby harnessing the substantial research value embedded within the dataset while upholding ethical considerations and societal well-being.

11. More on face recognition task

11.1. Face identification

Besides face verification, our dataset can also be used for face identification evaluation. Since we have multiple frames and multiple videos for an identity, we decided to develop both open-set and closed-set scenarios for the new benchmarking dataset. In the closed-set, we used 10,000 gallery items (1 per ID) and 50,000 probes. For open-set, 1,000 gallery items (1 per ID) and 5,000 mate searches, plus 32,275 non-mate searches from other IDs. For filtering and augmentation, we apply the same strategy as when building the face verification benchmark dataset. We followed NIST guidelines for metrics: FNIR at FPIR for open-set and TPIR at Ranks for closed-set. We tested Arcface Resnet-18/50 pretrained on Glint360K, Tab. 7 shows that pose remains a challenge in the face identification task.

11.2. Training face recognizer

For futher investigation of the pose distribution, we developed a training dataset of around 300,000 samples taken from identities that were not used for the proposed benchmark dataset. Due to time limitations, we trained only a ResNet-18 model using MS1MV3 along with this dataset for 20 epochs, following the same training settings from [7]. As shown if Tab. 8, our training dataset enhances the

	Open-set			Closed-set		
	FPIR	0.003	0.03	**Rank**	1	5
R18	**FNIR** ↓	0.45	0.26	**TPIR** ↑	0.80	0.96
R50	**FNIR** ↓	0.41	0.24	**TPIR** ↑	0.83	0.97

Table 7. **Quantitative result** of state-of-the-art models upon our proposed face identification benchmark dataset. We provided both scenarios: open-set and closed-set.

Training	EFHQ			CP	CF	IJBC
	f2f	f2p	p2p	LFW	PFP	
MS1MV3	0.93	0.57	0.88	0.75	0.82	0.94
+ EFHQ	**0.98**	**0.77**	**0.94**	**0.79**	**0.90**	0.94

Table 8. Effect of **EFHQ as additional training data** on model performance, with TAR@FAR=1e-4 for IJB-C and TAR@FAR=1e-3 for others.

trained model's performance on pose-focused benchmarks, albeit not marginally significant.

12. Training details

12.1. Face Reenactment

For face reenactment, we retrained two current state-of-the-art models, TPS and LIA, starting from scratch. This retraining was conducted on two datasets, including VoxCeleb1 and VoxCeleb1 supplemented with our EFHQ dataset. All models were trained using 2 Nvidia A100 GPUs with 40GB of memory. For both TPS and LIA models, we followed the hyperparameter choices outlined in the respective papers [43, 51].

In the case of VoxCeleb1, we followed the download and preprocessing pipeline as described in [38]. However, due to copyright and regional restrictions, we encountered limitations in fully acquiring the original dataset. Specifically, out of the 3,442 videos available in the original data, we were only able to download approximately 3,000 videos. This constraint limited our capacity to match the performance of the released pretrained models. Consequently, we opted to train both models from scratch for both datasets to ensure a fair benchmark for evaluation.

Thin-Plate Spline Motion Model. We set the total batch size to 28. The training process consists of two phases: the Base model phase, spanning 100 epochs with a learning rate of 0.0002, followed by the AVD Network phase, which extends for 200 epochs with a learning rate of 0.001. We utilize $K = 10$ TPS transformations to approximate the motion. In our loss function, we incorporate $\lambda = 10$ for the perceptual loss.

LIA. For LIA, we train both models with a total batch size of 32 and a learning rate of 0.002. The dimensions of all latent codes and the directions in the set of motion directions D_m are set to be 512. Meanwhile, we set the number of motion directions D_m to 20. In our loss function, we use $\lambda = 10$ for the perceptual loss.

12.2. ControlNet

Throughout both the training and sampling stages, a standardized set of prompt templates was consistently employed across all samples, formulated as "*A profile portrait image of a [emotion] [race] [gender].*" Specifically, in the sampling phase for evaluation, we implemented an augmentation strategy by integrating positive attributions, encompassing descriptors like "*rim lighting, studio lighting, dslr, ultra quality, sharp focus, tack sharp, depth of field (dof), film grain, Fujifilm XT3, crystal clear, 8K UHD, highly detailed glossy eyes, high detailed skin, and skin pores.*" Complementing these positive descriptors, we incorporated a set of negative keywords, including "*disfigured, ugly, bad, immature, cartoon, anime, 3d, painting, and black and white.*" The intent behind this approach was to refine the image quality within the sampling and evaluation process. This strategic augmentation framework aimed to enhance the overall quality and fidelity of generated images, ensuring a more refined output aligning with desired criteria and minimizing undesirable attributes.

13. CPLFW artifacts

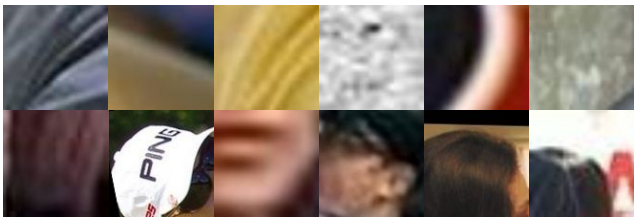

Figure 13. Representative examples of the noise sample due to misalignments in the widely-used processed version of CPLFW.

During the examination of the commonly utilized processed version of CPLFW[52] by [7], numerous instances of noise were identified, encompassing non-human and partial-human images. These anomalies likely arose from misalignments attributed to flawed landmark detections. To systematically address the bulk of these artifacts, we integrated a pretrained RetinaFace model [36] to automatically filter out images lacking detections. This filtering process resulted in the removal of 696 images, equating to 696 pairs, constituting approximately 11% of the total pairs. Fig. 13 showcases representative samples obtained through our cleaning pipeline. In the first row, we compile non-human images, whereas in the second row, we compile partial-human images.

14. More on the robustness of data processing pipeline.

To evaluate the robustness of our data processing pipeline, we test it with the raw LPFF dataset. As shown in Fig. 10, our pipeline detected additional profile-view faces filtered out by the LPFF pipeline, likely due to misdetections or overly strict pose density gating. The LPFF pipeline relies on three main processes for filtering: landmark prediction with Dlib[19], pose density gating using 3D face reconstruction[8], and manual inspection. The latter two stages depend heavily on accurate landmark predictions in the first step. Compared to Dlib, methods like RetinaFace[36] and SynergyNet[46] used in our pipeline offer improved face and landmark detection, especially for challenging poses. Our pipeline delivered markedly much superior landmark quality. To filter frontal samples, the previous work relied solely on the reconstruction model for pose estimation. Although [8] can provide state-of-the-art results, it depends on properly aligned images and may produce unreliable predictions when landmark quality is poor. Our pipeline uses an ensemble of pose estimators based on different methods, providing robustness in cases where a single model fails. By combining multiple pose prediction techniques, our approach can overcome the limitations of relying on just the output of reconstruction models, especially when image alignment is imperfect or landmarks are of low quality. This demonstrates the greater robustness of our approach to diverse poses and image variability compared to prior work. Moving forward, we hope our data processing methodology will enable the community to develop larger and higher-quality in-the-wild facial datasets. Robust pipelines to handle pose diversity and image noise will be essential to maximize data utilization and quality in these days and age.

15. Survey setup

15.1. 2D GAN Generation

This survey involves a comparative analysis between samples generated by StyleGAN2-ADA models trained with our setup (FFHQ+EFHQ) versus LPFF setup (FFHQ+LPFF). To ensure fairness, we specifically select pairs of samples exhibiting similar pose angles. Our assessment centers on two primary factors: Authenticity and Sharpness. In evaluating authenticity, participants are tasked with identifying the image with superior human face quality, considering artificial elements like unrealistic human features, skin tone, face shape, and facial accessories. Regarding sharpness, participants are prompted to examine image noisiness and select the image with generally higher image quality. In cases where a clear preference cannot be determined, participants have the option to indicate a tie. A total of 140 participants were engaged, contributing a combined total of 2800 votes over 20 questions.

15.2. EG3D

In this survey, we conduct a comparison between samples generated by EG3D models trained using our setup (FFHQ+EFHQ) and LFPP setup (FFHQ+LPFF). To ensure equitable assessment, we deliberately pair samples with similar gender, expression, and accessories whenever feasible. Each sample is presented through multiple views, including a frontal view for reference and four extreme views for comprehensive evaluation. Our analysis centers on three primary criteria: Pose, Identity, and Sharpness. Regarding pose assessment, participants are tasked with identifying the sample with superior face quality among the extreme views, specifically focusing on facial shape, such as nose and chin, while considering the presence of artifacts like unnatural distortions. In evaluating identity, participants are prompted to gauge how closely the facial identity aligns with the frontal view. Lastly, participants are asked to assess image quality for sharpness, emphasizing less noise and sharper details in the facial region. A total of 145 participants were engaged, contributing a combined total of 2900 votes, over 20 questions, to the study.

15.3. Face Reenactment

This study conducts a comparative analysis of models trained on both VoxCeleb1 and VoxCeleb1+EFHQ datasets for TPS and LIA [43, 51]. In order to offer a more comprehensive evaluation scenario for users, we compare video output sequences of same-identity reenactment between the baseline model and the model trained on our dataset rather than assessing individual frames. This approach facilitates a nuanced evaluation, especially in capturing the smoothness of pose transitions and illustrating video consistency when handling long sequences of extreme head poses. Our assessment focuses primarily on two key criteria: Identity and Motion. In the context of identity replication, our focus is on evaluating which model more accurately replicates the identity portrayed in the ground-truth driving video. This assessment takes into consideration facial features and the presence of any artificial artifacts. For motion replication, the focus shifts to discerning which model more accurately reproduces the sequence of motion represented in the driving video. In cases where users find it challenging to determine superiority, they have the option to indicate a tie. In total, we involved 170 users, who collectively contributed 8500 votes to this evaluation, addressing 25 questions for each face reenactment method.

16. Additional Qualitative results

We provide more qualitative examples of 2D/3D-aware GAN generation, diffusion-based text-to-image generation,

and face reenactment to further demonstrate the superiority of our method. The short descriptions for the figures are shown below.

- Fig. 14 shows a comparative analysis conducted at the patch level between samples from StyleGAN2-ADA trained with FFHQ+LPFF and FFHQ+EFHQ datasets to further elucidate the quality and sharpness of the synthesized facial images. Fig. 15 exhibits additional samples generated by our models.
- Figs. 16 to 20 present a comparison of synthesized faces from various views produced by EG3D, trained with FFHQ, FFHQ+LPFF, and FFHQ+EFHQ datasets. Moreover, Figs. 21 and 22 serve the same purpose as the aforementioned figures but with a different setup of viewing angles.
- Fig. 23 present a comparison of synthesized faces from ControlNet trained with OpenPose[4] 's dataset (released by ControlNet[49]) and trained with our dataset.
- Figs. 24 to 26 present a comparison of same-identity face reenactment from TPS, trained with VoxCeleb1 and with VoxCeleb1+EFHQ.
- Figs. 27 to 29 present a comparison of same-identity face reenactment from LIA, trained with VoxCeleb1 and with VoxCeleb1+EFHQ.
- Fig. 30 presents a comparison of cross-identity face reenactment from both LIA and TPS, trained with VoxCeleb1 and with VoxCeleb1+EFHQ.

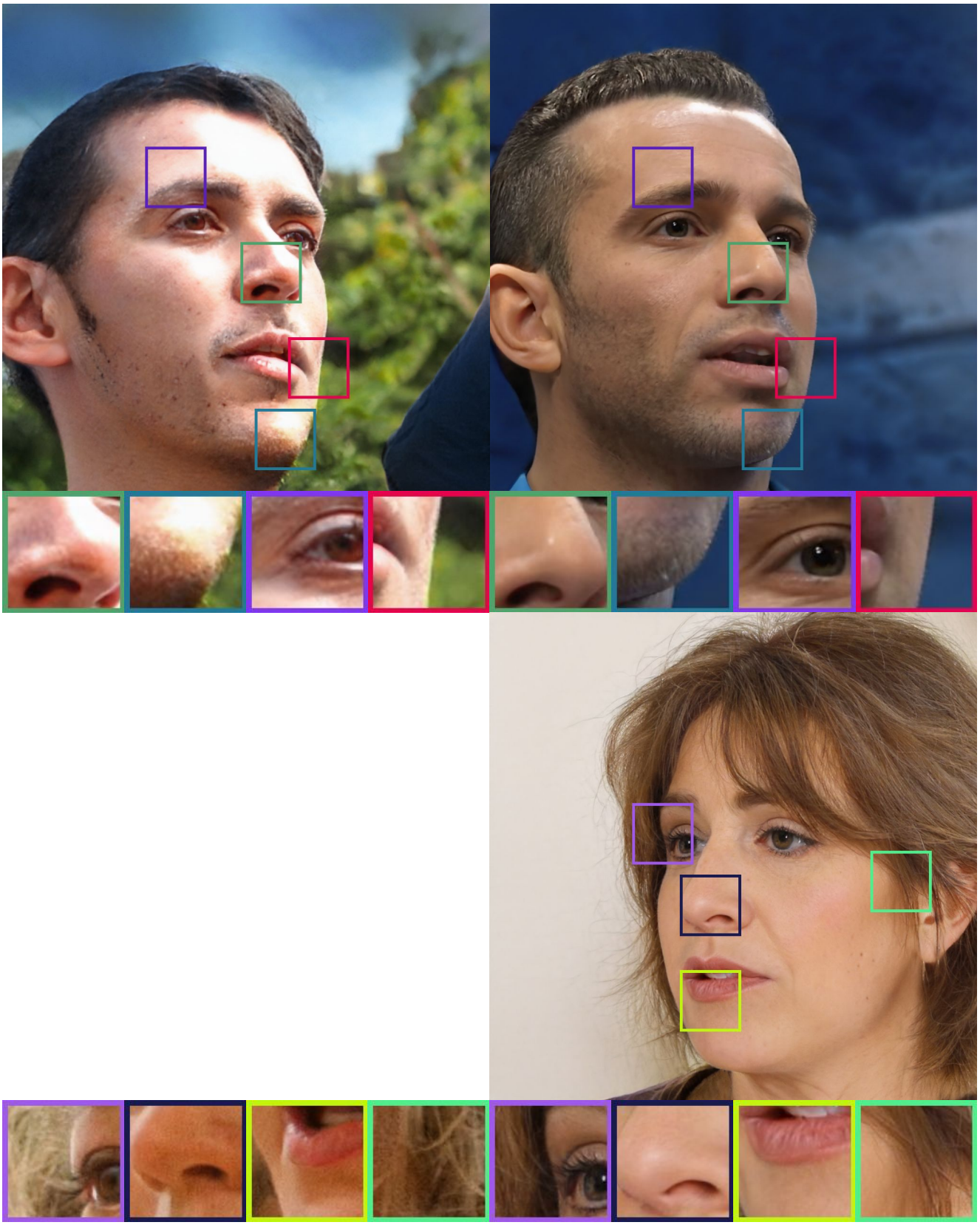

Figure 14. Comparison between profile-view generated samples of StyleGAN2-ADA training with FFHQ+LPFF (left) and FFHQ+EFHQ (right), with truncation $\psi = 0.7$.

Figure 15. Samples from the model trained with FFHQ+EFHQ, with truncation $\psi = 0.7$..

Figure 16. **Comparison** between multiview generated samples, with truncation $\psi = 0.8$, of EG3D model trained with various datasets (for both LPFF/EFHQ, the training dataset is combined with FFHQ).

Figure 17. **Comparison** between multiview generated samples, with truncation $\psi = 0.8$, of EG3D model trained with various datasets (for both LPFF/EFHQ, the training dataset is combined with FFHQ).

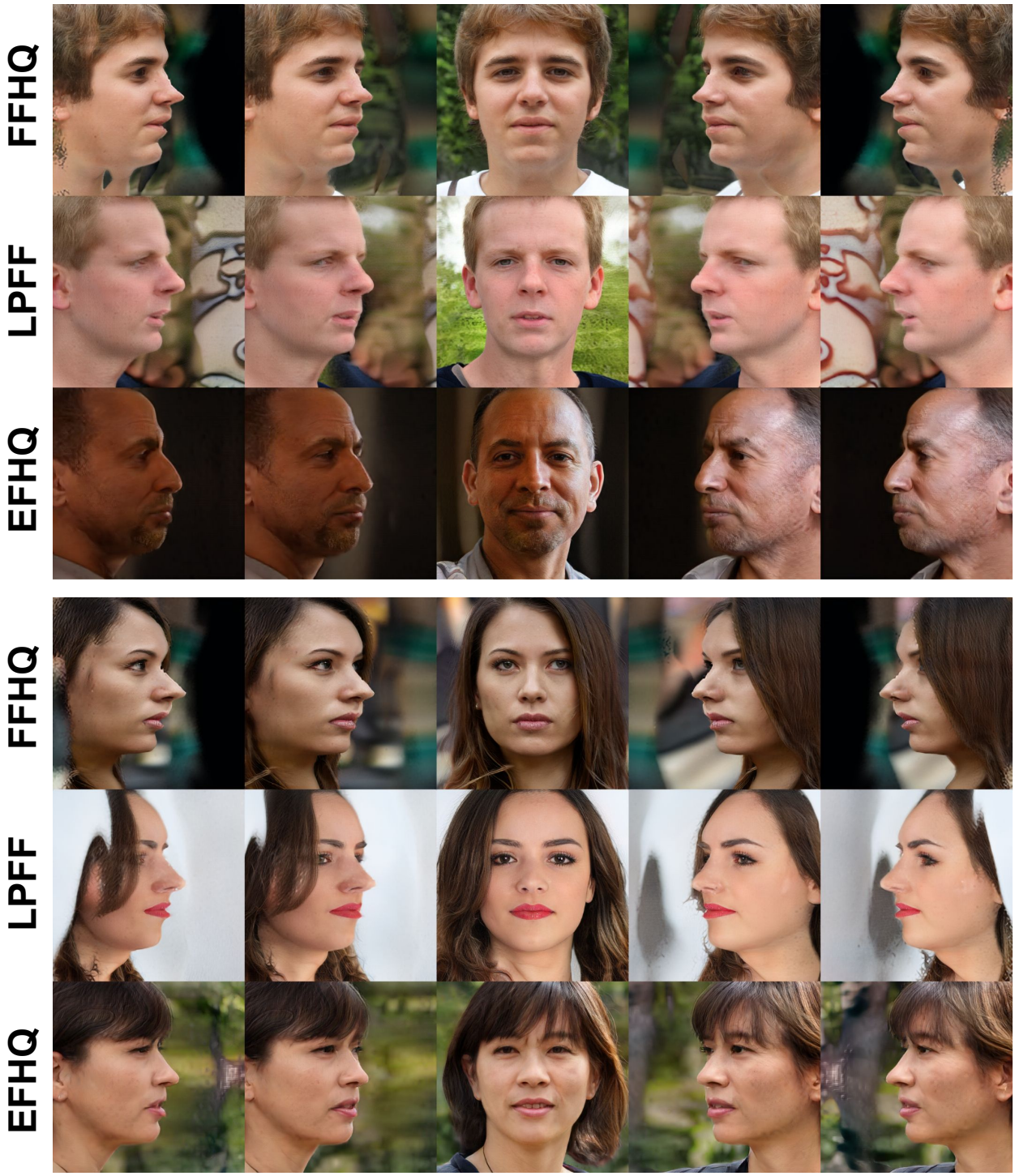

Figure 18. **Comparison** between multiview generated samples, with truncation $\psi = 0.8$, of EG3D model trained with various datasets (for both LPFF/EFHQ, the training dataset is combined with FFHQ).

Figure 19. **Comparison** between multiview generated samples, with truncation $\psi = 0.8$, of EG3D model trained with various datasets (for both LPFF/EFHQ, the training dataset is combined with FFHQ).

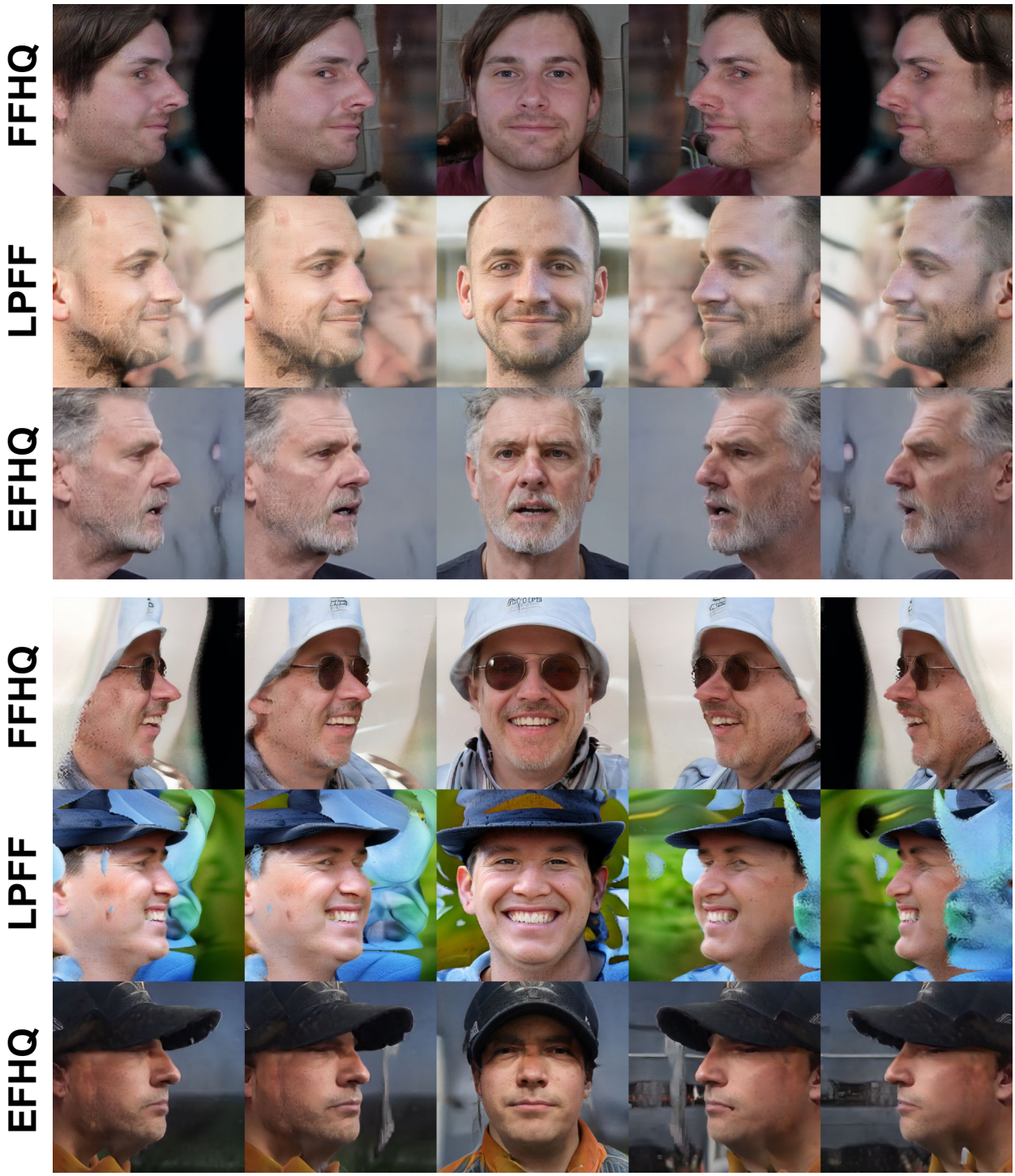

Figure 20. **Comparison** between multiview generated samples, with truncation $\psi = 0.8$, of EG3D model trained with various datasets (for both LPFF/EFHQ, the training dataset is combined with FFHQ).

Figure 21. **Comparison** between multiview generated samples, with truncation $\psi = 0.8$, of EG3D model trained with various datasets (for both LPFF/EFHQ, the training dataset is combined with FFHQ).

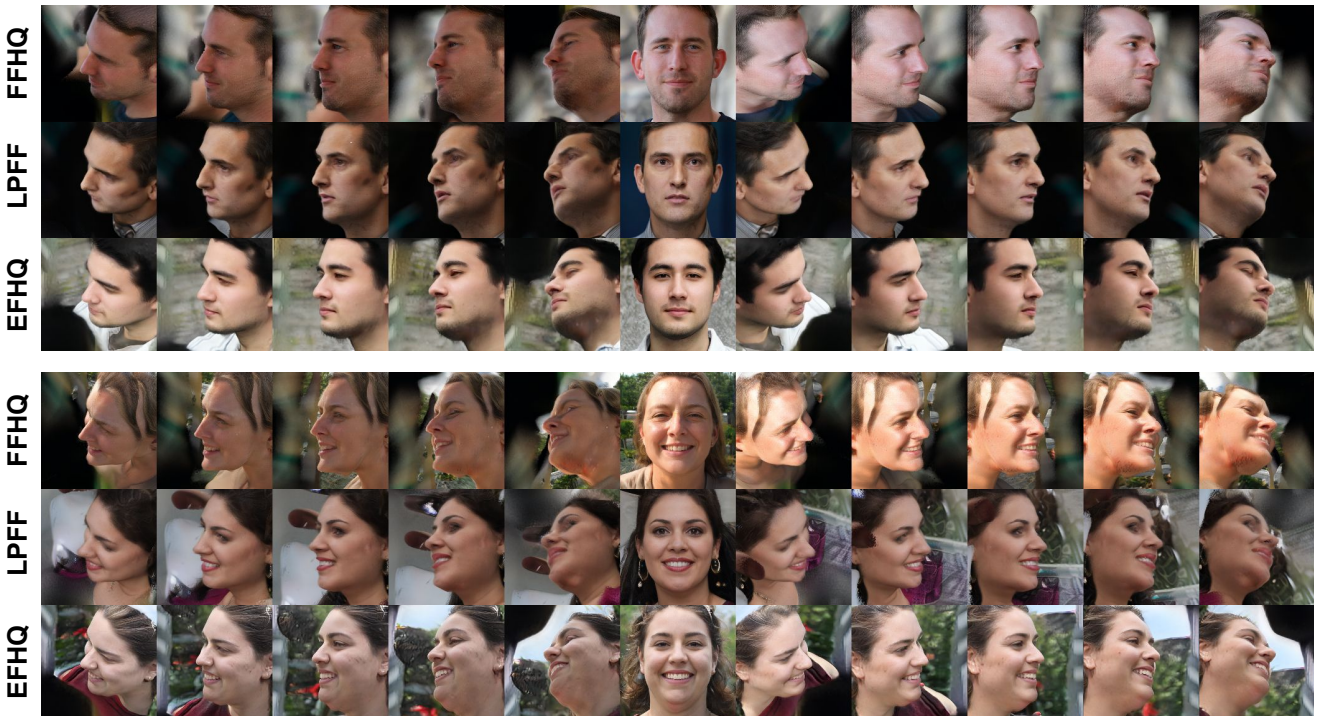

Figure 22. **Comparison** between multiview generated samples, with truncation $\psi = 0.8$, of EG3D model trained with various datasets (for both LPFF/EFHQ, the training dataset is combined with FFHQ).

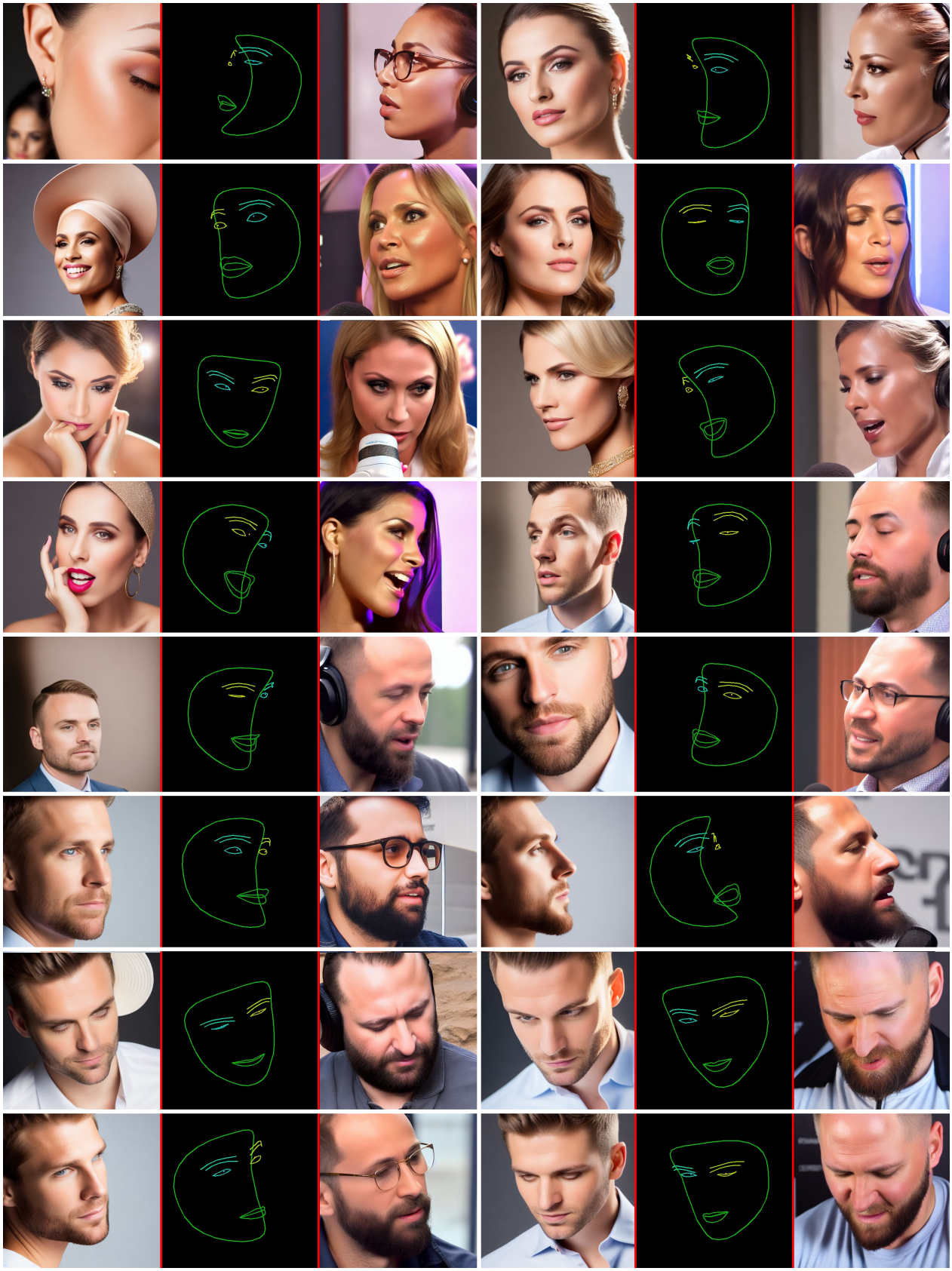

Figure 23. **Comparison** profile-view generated samples of pretrained ControlNet (left) and our fine-tuned ControlNet (right) with the prompt: "A profile portrait image of a person."

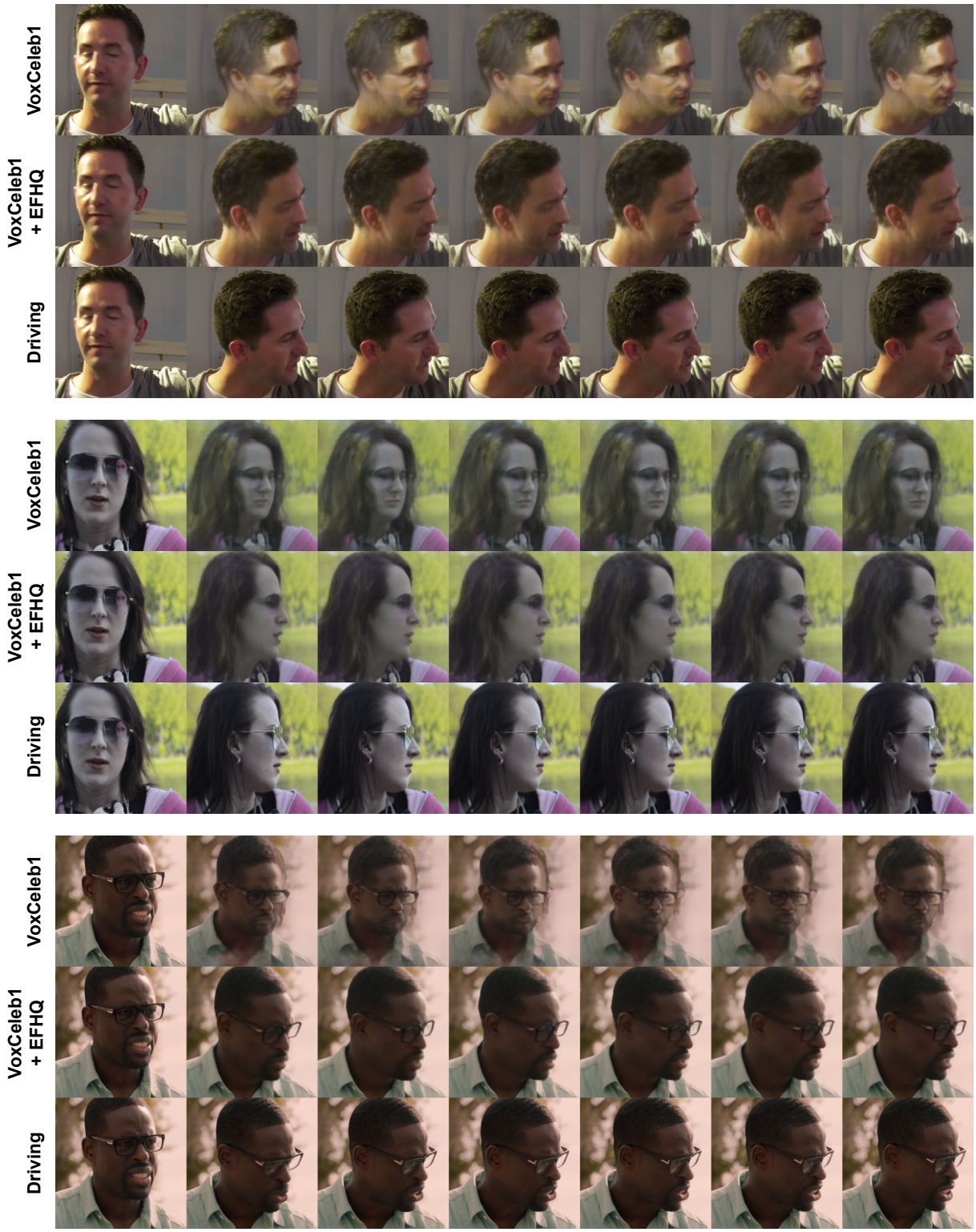

Figure 24. **Comparison** between same-identity reenactment of TPS model trained with various datasets. The first frame of each row represents the source image, while the last row depicts the ground truth driving frames.

Figure 25. **Comparison** between same-identity reenactment of TPS model trained with various datasets. The first frame of each row represents the source image, while the last row depicts the ground truth driving frames.

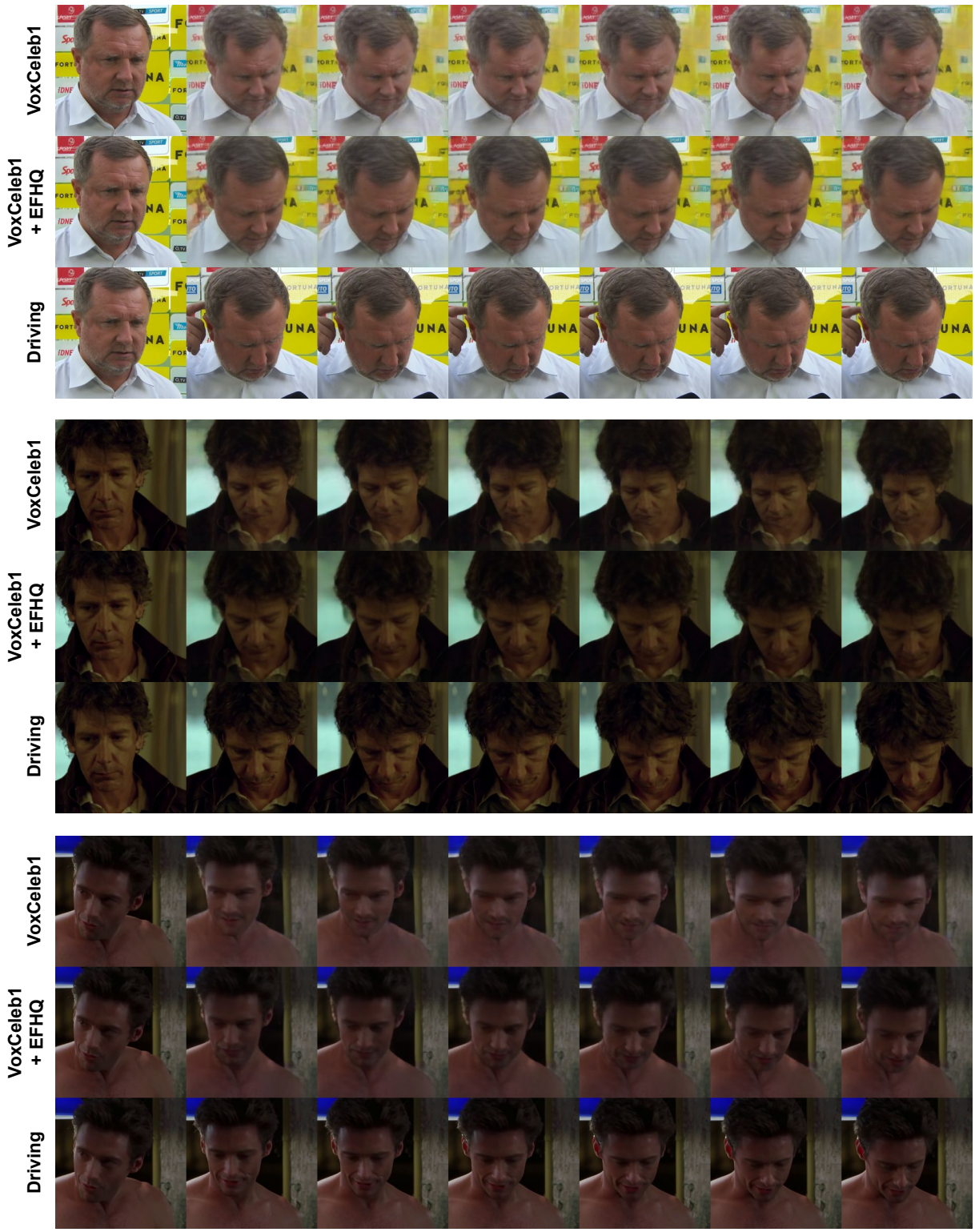

Figure 26. **Comparison** between same-identity reenactment of TPS model trained with various datasets. The first frame of each row represents the source image, while the last row depicts the ground truth driving frames.

Figure 27. **Comparison** between same-identity reenactment of LIA model trained with various datasets. The first frame of each row represents the source image, while the last row depicts the ground truth driving frames.

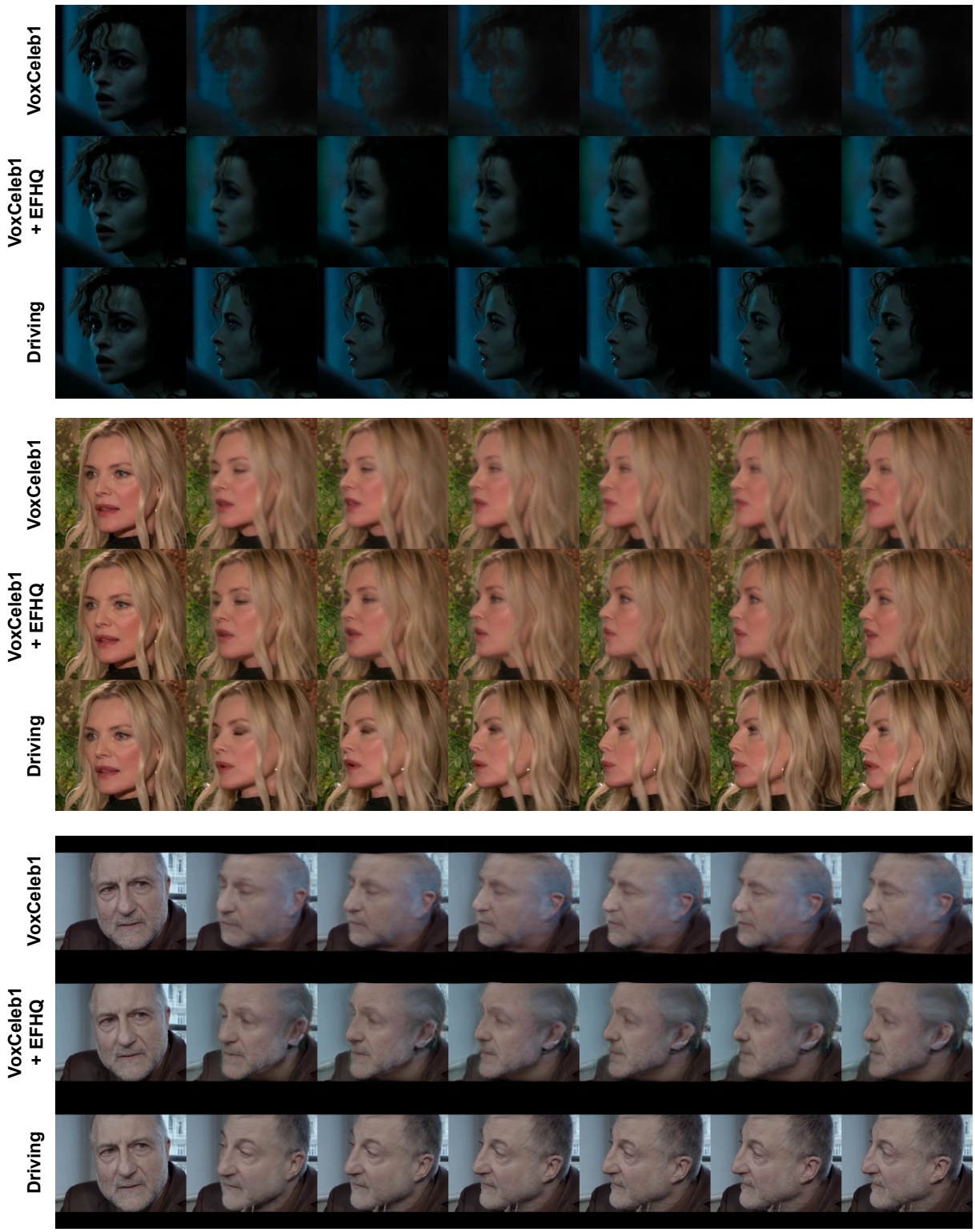

Figure 28. **Comparison** between same-identity reenactment of LIA model trained with various datasets. The first frame of each row represents the source image, while the last row depicts the ground truth driving frames.

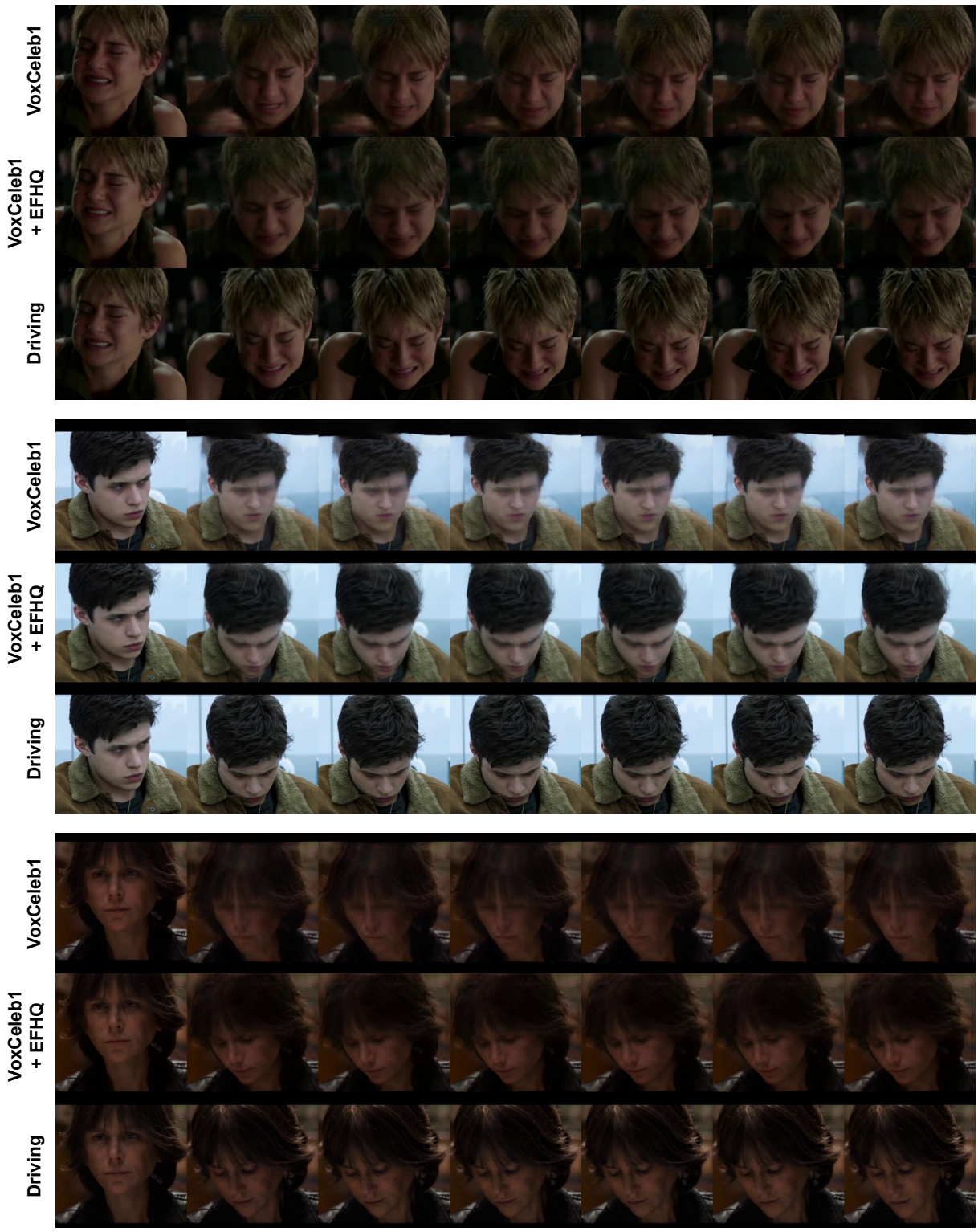

Figure 29. **Comparison** between same-identity reenactment of LIA model trained with various datasets. The first frame of each row represents the source image, while the last row depicts the ground truth driving frames.

Figure 30. **Comparison** between cross-identity reenactment of models trained with various datasets. The first frame of each row represents the source image, while the last frame depicts the ground truth driving frame.