Anchored Diffusion for Video Face Reenactment

Supplementary Material

1. Limitations

While our method shows strong potential for generating videos, its current demonstration is limited to face reenactment using facial landmarks as the primary control signal. One observed limitation is the preservation of head shape attributes from the driving video, which can hinder precise identity preservation. Future work could explore several promising directions to address this and other limitations. First, we can expand the types of guidance beyond facial landmarks, incorporating segmentation maps, optical flow, depth maps, and other modalities. Second, we can apply our approach to diverse domains like image-to-video, inverse problems, cinemagraphs, and special effects. To mitigate head shape preservation, we plan to investigate mapping landmarks to a canonical space and adjusting them based on differences with the source image's landmarks. Finally, our method is currently limited to a single scene or shot, but we envision using multiple or moving anchors to enable multi-scene video generation.

2. Social Impact

We recognize the risks associated with our large-scale human-face dataset sourced from YouTube, including but not limited to privacy, copyright violations, potential misuse, and inherent biases. To mitigate these concerns and potential harm, we implement strict access control measures and a multi-faceted approach. Importantly, we only publish links to public videos, which helps protect privacy, reduces legal liability, and facilitates an easy opt-out mechanism. Furthermore, we maintain strict control over dataset access by enforcing a research-only license and limiting distribution to authorized researchers. We are committed to transparency by publishing reports detailing our dataset's composition, usage statistics, opt-out requests, and any detected misuse attempts. We strive to balance research advancement with ethical considerations, remaining adaptable to evolving challenges in this domain.

3. Artistic Reenactment

We introduce an additional application: artistic reenactment, which involves transferring facial expressions and movements from a driving video to a target artistic portrait. To address the domain gap between our curated training data clips and the desired artistic domain, we incorporate 25,000 artistic images from the Artstation-Artistic-face-HQ (AAHQ) [5] dataset into our training scheme. Geometric transformations are applied to each sample to conceptualize a series of video clips.

Interestingly, as demonstrated in Fig. 1 and Fig. 2, although the geometric transformations do not include detailed movement information, such as different poses or eye closure, the model successfully learns to reenact artistic video clips. This success is attributed to the integration with real-world clips, which enables the model to effectively bridge the gap between artistic and realistic domains.

4. Ablation Study

In this section, we conduct ablation studies using various inference configurations for cross-identity reenactment on our Records-Test-5K dataset. We evaluate the realism of the generated outputs using the Fréchet Inception Distance (FID) [2] and assess motion transfer accuracy by extracting 478 XYZ facial landmarks from both the generated and driving videos using MediaPipe [7]. The accuracy is quantified by calculating the Mean Squared Error (MSE) between corresponding points, denoted as LMSE. To measure scene consistency throughout the video, we compute the minimum cosine similarity (CSIM) between the source and generated embeddings using CLIP.

Table 1 summarizes the quantitative results across different inference configurations. First, we examine the impact of classifier-free guidance (CFG), which controls the influence of guidance on image generation. Interestingly, disabling CFG in our model leads to better performance. We then evaluate the effect of varying the number of timesteps during generation, confirming that a sampling procedure with 1000 steps yields the best results. Finally, we investigate the role of guidance during the sampling stage. Since the model is trained with random condition dropout (with a 10% probability) to support CFG, inference without any guidance can generate realistic images, though they lack correlation to the driving video or source image. When using only landmarks or only CLIP guidance, the model performs poorly, as it was not trained for these scenarios. The suggested configuration in the last row of the table achieves

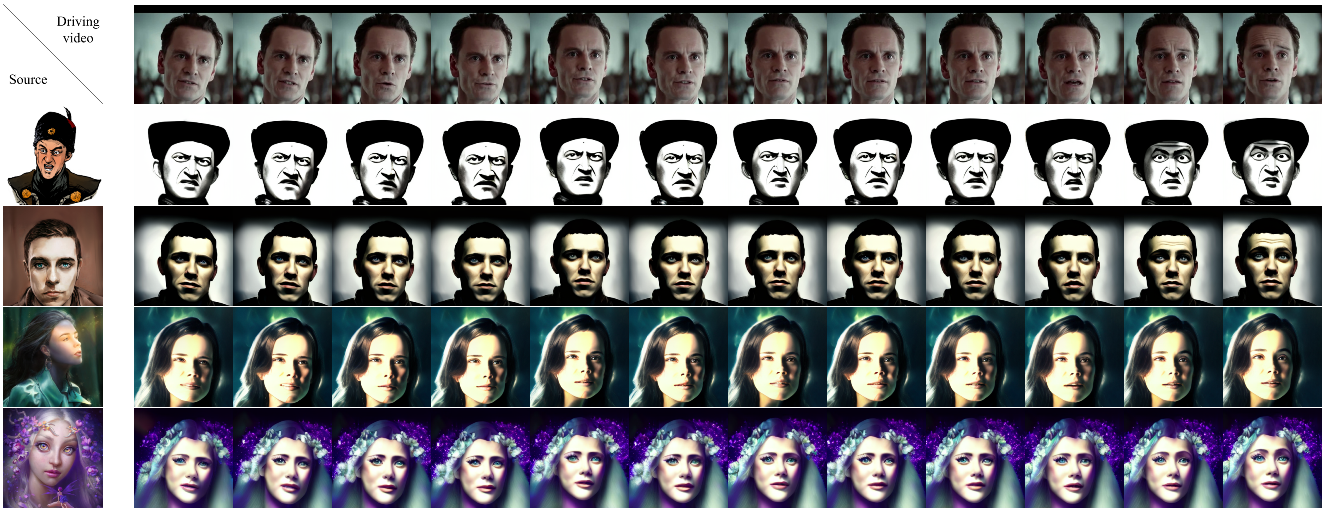


Figure 1. **Artistic Reenactment.** Results of the artistic reenactment process.

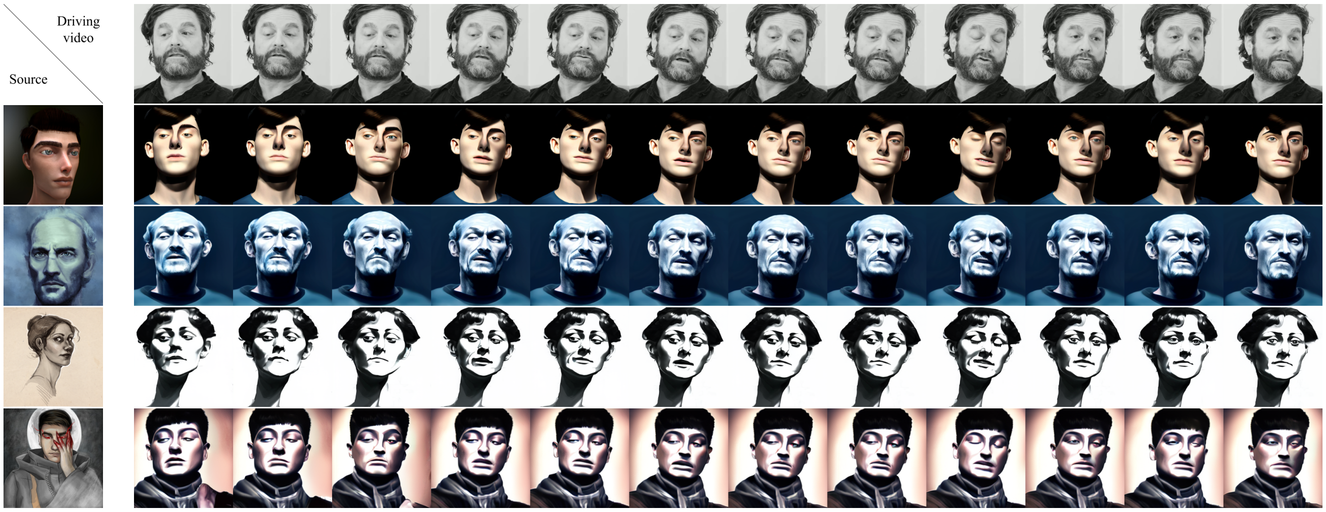


Figure 2. **Artistic Reenactment.** Additional results of the artistic reenactment process.

the best overall performance across all metrics.

5. Dataset Statistics

In this work, we introduce *ReenactFaces-1M*, a large-scale, high-quality, and diverse video dataset. *ReenactFaces-1M* comprises $1,006,257$ video segments, each with an average length of 3.29 seconds, totaling over 920 hours of footage. The dataset exhibits an average resolution of 745 pixels, making it a valuable resource for various applications in video analysis and facial recognition research. To further understand the characteristics of our dataset, we provide an analysis of important data statistics:

- Figure 3 shows the distribution of clip durations in our dataset, with an average duration of 3.29 seconds and a standard deviation of 2.07 seconds.
- Figure 4 shows the distribution of clip HyperIQA [12] scores in our dataset, with an average duration of 51.5 and a standard deviation of 10.72.
- Figure 5 shows the distribution of clip resolution in our dataset, with an average duration of 745.1 and a standard deviation of 247.8.
- Figure 6 depicts the distribution of the face height ratio relative to the total clip height and the face width ratio relative to the total clip width. The face width ratio has a mean of 0.45 and a standard deviation of 0.05, while the face height ratio has a mean of 0.53 and a standard deviation of 0.07.

Timesteps	CFG	Guidance	FID	LMSE	CSIM
1000	1	CLIP+Landmarks	34.4	9.61	0.74
1000	2	CLIP+Landmarks	34.6	10.31	0.77
1000	4	CLIP+Landmarks	38.5	12.63	0.79
1000	8	CLIP+Landmarks	48.0	14.14	0.77
50	1	CLIP+Landmarks	36.6	12.34	0.70
100	1	CLIP+Landmarks	34.0	11.21	0.72
200	1	CLIP+Landmarks	34.3	10.82	0.73
400	1	CLIP+Landmarks	34.2	9.93	0.72
1000	1	CLIP+Landmarks	34.4	9.61	0.74
1000	1	None	67.8	582	0.46
1000	1	Landmarks	361	3268	0.43
1000	1	CLIP	287	1992	0.50
1000	1	CLIP+Landmarks	34.4	9.61	0.74

Table 1. **Ablation Quantitative Results.** Comparisons with various inference configurations on the cross-identity reenactment using our Records-Test-5K dataset.

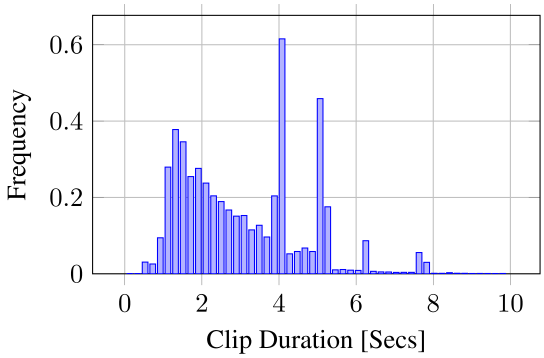


Figure 3. **Clip Duration.** This histogram shows the distribution of clip durations in our dataset, with an average duration of 3.29 seconds and a standard deviation of 2.07 seconds.

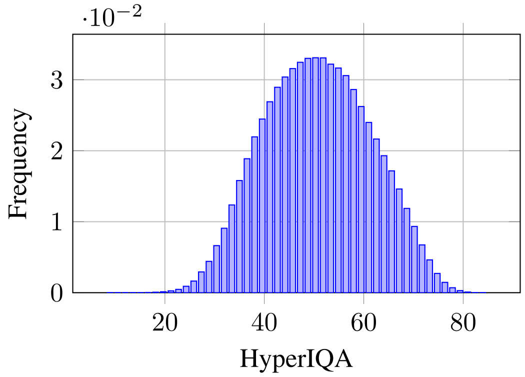


Figure 4. **Clip HyperIQA.** This histogram shows the distribution of clip HyperIQA scores in our dataset, with an average duration of 51.5 and a standard deviation of 10.72.

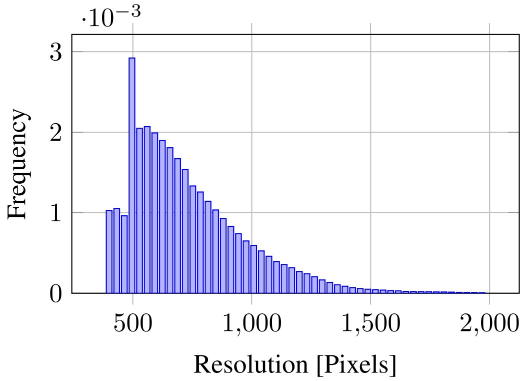


Figure 5. **Clip Resolution.** This histogram shows the distribution of clip resolution in our dataset, with an average duration of 745.1 and a standard deviation of 247.8.

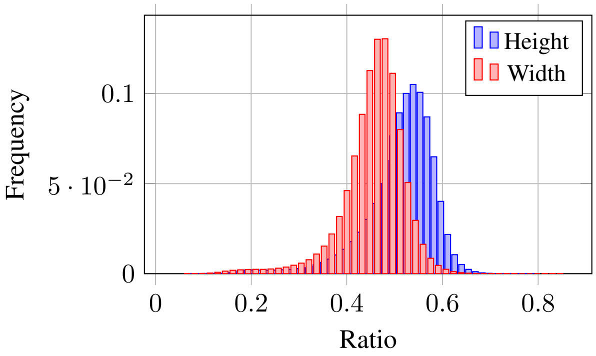


Figure 6. **Facial Ratio.** The histogram depicts the distribution of the face height ratio relative to the total clip height and the face width ratio relative to the total clip width. The face width ratio has a mean of 0.45 and a standard deviation of 0.05, while the face height ratio has a mean of 0.53 and a standard deviation of 0.07.

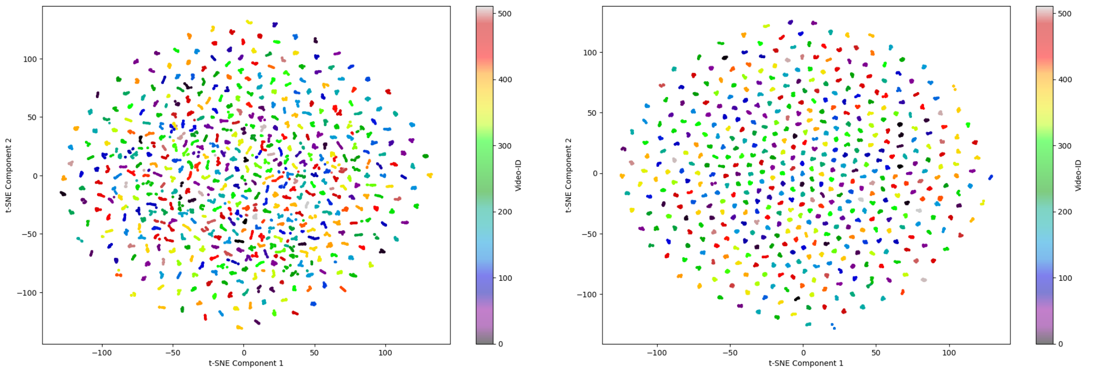
6. Face Reenactment - Extended Results

6.1. Analysis of Scene Recognition

To assess the scene recognition capabilities of our approach, we analyzed both ArcFace and CLIP embeddings of video frames. Figure 7 presents t-SNE visualizations of these embeddings, where each point represents a frame and its color corresponds to the video it belongs to. The ArcFace embeddings are not as well-separated, failing to distinguish between certain videos. In contrast, the CLIP embeddings resulted in clear separation of clusters, indicating that they effectively distinguish between different scenes and movies, highlighting their potential for such tasks.

6.2. Additional Visual Results

To further illustrate the capabilities of our face reenactment approach, we present a broader range of visual results in Figures 8-12. These examples highlight the model's abil-



ArcFace

CLIP

Figure 7. **Scene Recognition.** t-SNE 2D projection of ArcFace Embeddings (Left) and CLIP embeddings (Right).

ity to handle challenging conditions such as extreme poses and varying facial attributes, while maintaining visual fidelity and temporal consistency. Additionally, Figure 13 highlights the model's effectiveness in generating coherent and extended video sequences, further demonstrating its versatility and potential applications.

7. Additional Experimental Details

We train two base sDiT-XL models at a resolution of 256x256 pixels, each with a patch size of 2x2. These models are capable of generating sequences of 4 and 8 frames, respectively. The mapping network consists of 4 residual blocks, and we use standard weight initialization techniques from ViT [1]. All models are trained with AdamW [6], using default parameter values and a cosine learning rate scheduler. The initial learning rates are set to 6.4×10^{-5} for the denoiser and 6.4×10^{-6} for the mapping network.

For the VAE model, we use a pre-trained model from Stable Diffusion [10]. The VAE encoder downscales the spatial dimensions by a factor of x8 while producing a 4-channel output for a 3-channel RGB input. We retain diffusion hyperparameters from DiT [9], including $t_{max} = 1000$ and a learned sigma routine.

Our training loss function is a weighted MSE, designed to prioritize accurate reconstruction of facial expressions, specifically targeting facial landmarks around the mouth and eyes. These expressive landmark pixels are assigned a weight of $(1 + \lambda_{\text{ex}})$, with λ_{ex} set to 1, while other pixels are weighted at 1.

All models are trained for 1 million steps using a global batch size of 16 samples. We implement our models in Pytorch [8] and train them using four Nvidia A100-SXM4-80GB GPUs. The most compute-intensive model achieves a training speed of approximately 1.8 iterations per second.

References

[1] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 4

[2] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 1

[3] 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. 5, 6, 7

[4] Fa-Ting Hong, Longhao Zhang, Li Shen, and Dan Xu. Depth-aware generative adversarial network for talking head video generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3397–3406, 2022. 5, 6, 7

[5] Mingcong Liu, Qiang Li, Zekui Qin, Guoxin Zhang, Pengfei Wan, and Wen Zheng. Blendgan: Implicitly gan blending for arbitrary stylized face generation. In *Advances in Neural Information Processing Systems*, 2021. 1

[6] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 4

[7] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, et al. Mediapipe: A framework for building perception pipelines. *arXiv preprint arXiv:1906.08172*, 2019. 1

[8] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019. 4

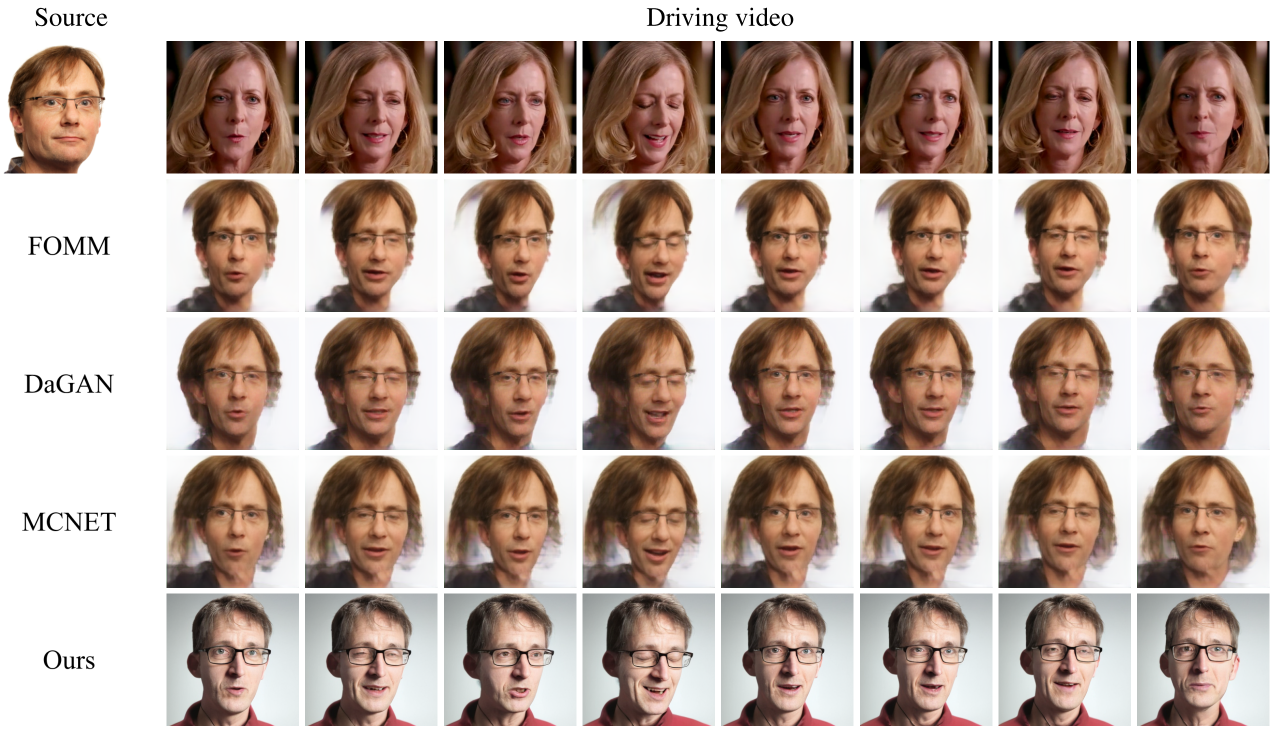


Figure 8. **Cross-identity Reenactment.** Comparisons with the competing methods [3, 4, 11].

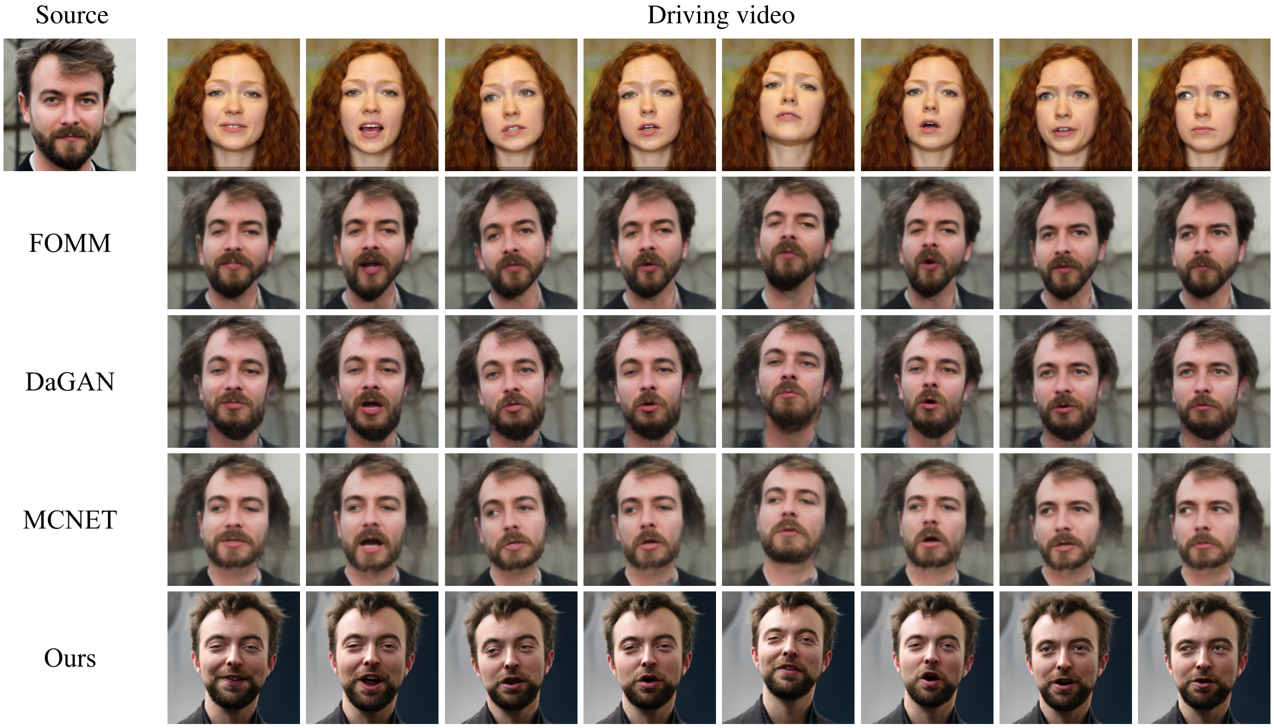


Figure 9. **Cross-identity Reenactment.** Comparisons with the competing methods [3, 4, 11].

[9] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205, 2023. 4

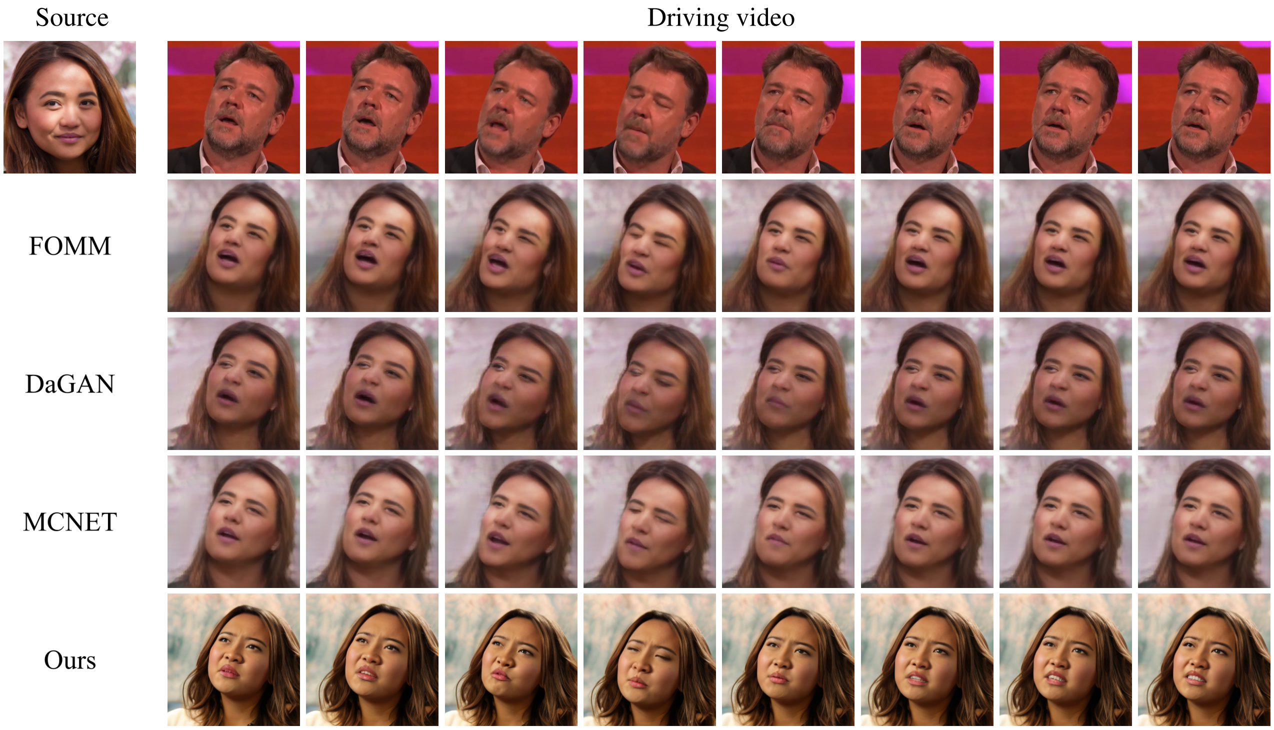


Figure 10. **Cross-identity Reenactment.** Comparisons with the competing methods [3, 4, 11].

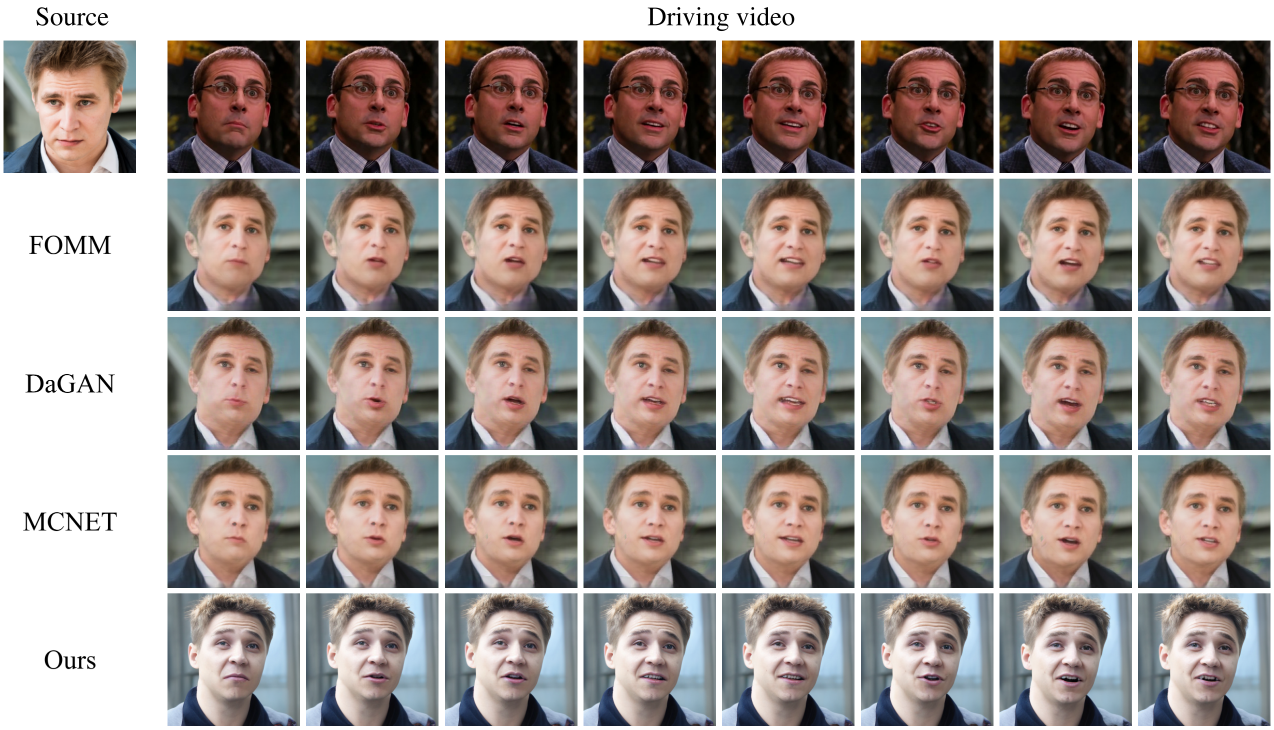


Figure 11. **Cross-identity Reenactment.** Comparisons with the competing methods [3, 4, 11].

[10] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern*

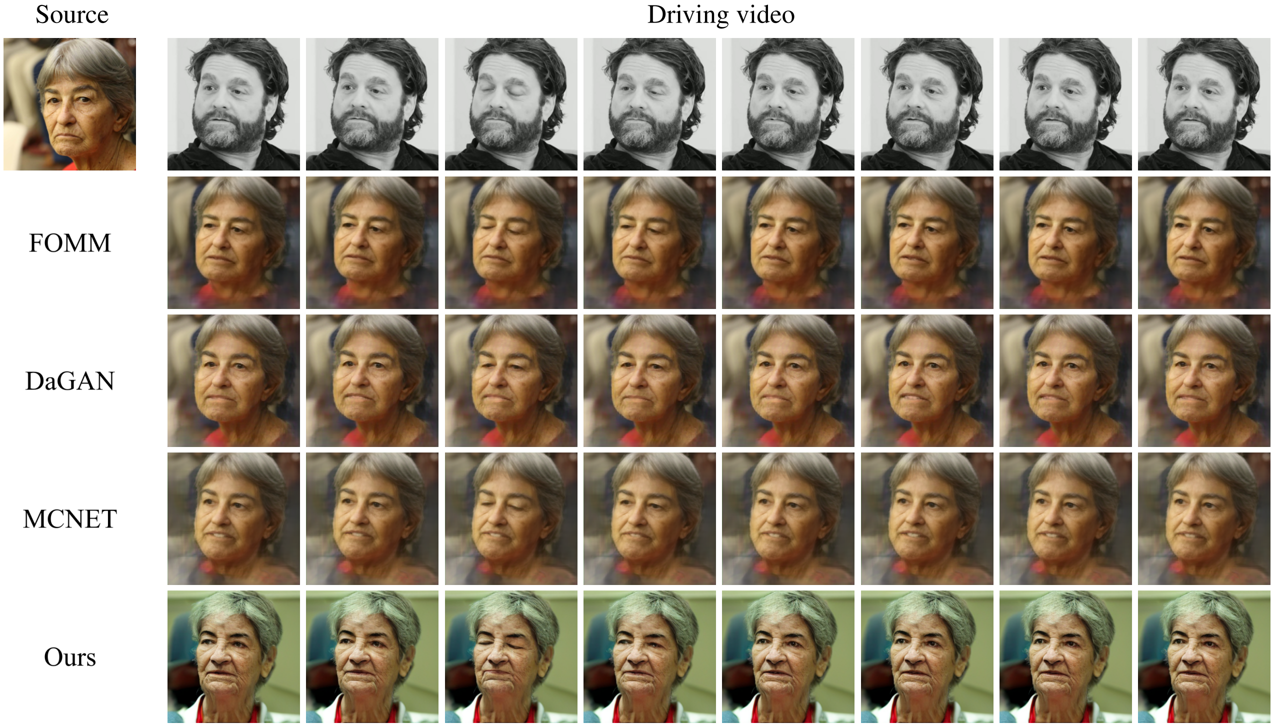


Figure 12. **Cross-identity Reenactment.** Comparisons with the competing methods [3,4,11].

recognition, pages 10684–10695, 2022. 4

[11] 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. 5, 6, 7

[12] Shaolin Su, Qingsen Yan, Yu Zhu, Cheng Zhang, Xin Ge, Jinqiu Sun, and Yanning Zhang. Blindly assess image quality in the wild guided by a self-adaptive hyper network. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3667–3676, 2020. 2

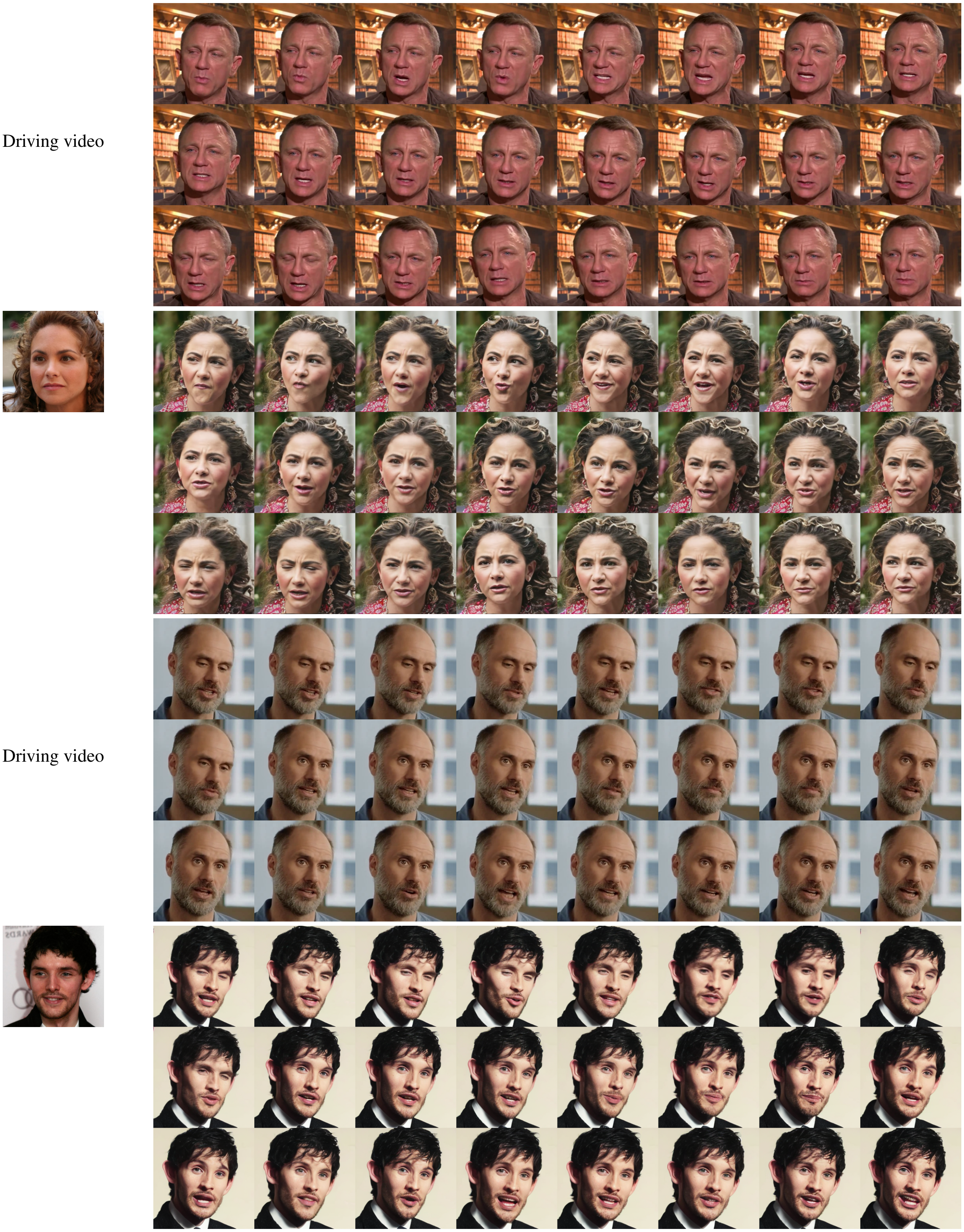


Figure 13. **Cross-identity Reenactment.** Enabling Identity-Swapping in 24-Frame Video Clips.