

# Joker: Conditional 3D Head Synthesis with Extreme Facial Expressions

## Supplementary Material



Figure 1. Further 3D reconstructions of our method on a diverse set of examples including out-of-distribution samples.

### 1. Datasets

As described in the main paper, we use the datasets CelebV-Text [25] and NeRSemble [12] and generate new annotations and metadata to train our model. In addition, we recorded challenging samples for validating the synthesis of extreme facial expressions in an in-the-wild setting; this benchmark is referred to as *Joker benchmark*.

**CelebV-Text Dataset** The CelebV-Text Dataset [25] is a large-scale facial text-video dataset containing 70k facial video clips from the internet with a total length of 279 hours. We first filter out videos of low quality by discarding samples with a HyperIQA score [18] of less than 40. Second, we filter for videos with extreme and diverse poses and expressions. For that, we use an off-the-shelf model<sup>1</sup> [6] to annotate the video frames with 3DMM parameters and select the frames with the highest expressiveness and diversity. The filtered images are cropped following the alignment procedure of [8] and automatically annotated with BFM parameters and text captions using Deep3DFaceRecon [8] and Blip2 [13]. Samples for which the 3DMM parameters estimation fails and with implausible captions are discarded.

We select 50k samples for training and 2.5k for evaluation. Reference images are randomly sampled from the same sequence as the target image, weighted by the relative

distance in pose and expression. To avoid identity overlap between the training and validation sets, we use an off-the-shelf face recognition network [7] and enforce an identity similarity score of less than 0.4 between each validation sample and its closest training sample. The automatically generated annotations and metadata will be made publicly available to the research community.

**NeRSemble Dataset** The NeRSemble dataset [12] is a multi-view portrait video dataset containing 4734 recordings of 222 subjects captured with 16 machine vision cameras. The subjects perform a wide set of extreme expressions in an environment with uniform lighting and background. We follow the same procedure as for CelebV-Text for sample filtering, image cropping, and annotation. Further, we assign a higher sampling ratio to the samples for which the automatically generated caption contains the keyword "tongue" because such samples are sparse in the CelebV-Text dataset. Note that we only create the image captions for the frontal images and reuse them for the other multi-view images. Reference images are randomly sampled from images showing the same subject as the target image but with a different expression and captured from a different camera. We split the dataset into 199 subjects for training and 23 for validation and automatically selected 2,000 and 2,500 frames, respectively.

<sup>1</sup>[https://github.com/radekd91/inferno/tree/master/inferno\\_apps/FaceReconstruction](https://github.com/radekd91/inferno/tree/master/inferno_apps/FaceReconstruction)



Figure 2. Random samples from our *Joker benchmark*. The samples contain in-the-wild scenes with natural backgrounds and lighting and studio scenes with uniform backgrounds and lighting.

**Joker Benchmark** Evaluating our method on the validation sets of CelebV-Text and NeRSemble alone is insufficient: CelebV-Text only contains moderately extreme expressions, and NeRSemble is restricted to uniform lighting and background scenarios. For this reason, we captured the *Joker benchmark* for the evaluation of extreme expression synthesis, which will be made publicly available to the research community. It provides monocular videos of 13 subjects performing extreme expressions both in in-the-wild scenarios, as well as in a lab environment with uniform lighting and background, see Figure 2. The subjects are of diverse ethnicity and equal gender parity (6 male, 7 female). We apply the same alignment and annotation pipeline to the dataset as for CelebV-Text.

## 2. Description of the baseline methods

*VOODOO 3D* [20] finetunes a pretrained model [21] to lift the reference image into 3D and trains a model to transfer expressions between the 3D representations of the driving and the reference subject. *VOODOO XP* [19] similarly to *VOODOO 3D* also leverages 3D lifting but learns an expression encoder in an end-to-end fashion to provide fine-grained expression control. *Real3D-Portrait* [24] combines an image-to-plane model with a tri-plane motion adapter to synthesize 3D talking head avatars that can be controlled via audio or 3DMM parameters. *Portrait4D-v2* [9] combines a modified EG3D [2] pipeline with a control mechanism through the FLAME 3DMM [14]. *GOHA* [15] uses a 3DMM to control facial expressions by mapping 3DMM parameters to residuals of a tri-plane representation of the face. *AniFaceDiff* [3] follows a similar approach as our method, yet instead of using a ControlNet, they encode

normal maps of FLAME [14] through stacked 2D convolutions and directly add them to the noisy input latents. Further, they don’t use text control but apply cross-attention to features extracted from the FLAME parameters. *X-Portrait* [23] also follows a similar approach as our method. In contrast to our method, however, they don’t utilize text and 3DMM as inputs. Instead, they use patches of the driving image as input to the ControlNet. To avoid identity leakage during training, *X-Portrait* uses a pre-trained facial reenactment method to generate them.

Note that for the baselines *Real3DPortrait*, *AniFaceDiff*, and *X-Portrait*, we use the renderings of our method to obtain the dynamic camera sweep results presented in the suppl. video. For the other baselines, we directly use the ground truth camera parameters for rendering.

## 3. Additional Experiments

### 3.1. Ablation Study of Our 2D Prior

We ablate the design choices of our 2D prior in Table 1 and Figure 3. In contrast to *X-Portrait*, we unfreeze the up-sampling blocks of our denoising UNet and find that this consistently improves all metrics. Qualitatively, we observe particularly significant improvements for identity preservation under extreme expression changes (see results ‘Frozen Denoising UNet’ in Figure 3).

Removing the text control from our method during training and inference significantly worsens all metrics (see results for ‘No Text Control’). Qualitatively, we observe that extreme expressions, most prominently tongue articulations, cannot be controlled through 3DMM parameters alone, which explains the observed deterioration of the evaluation scores. Note that none of the existing methods can



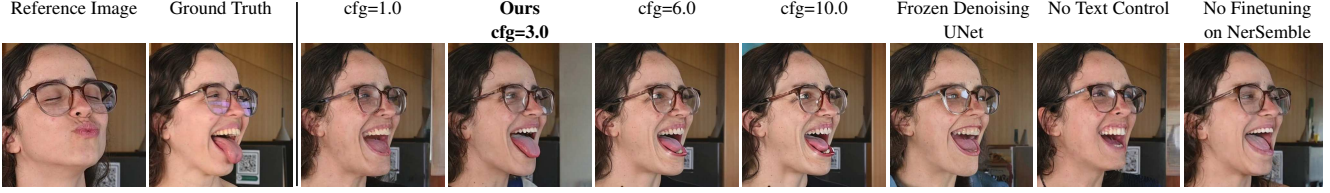


Figure 3. Qualitative ablation study of our **2D prior**. Too small classifier-free guidance scale values (cfg) reduce the faithfulness of extreme expressions, while too high values cause oversaturation and artifacts. We find that  $\text{cfg}=3$  yields the best trade-off. Not training the upsampling layers of the denoising UNet ("*Frozen Denoising UNet*") worsens identity preservation and synthesis quality in general. Dropping the text control disables tongue control since it is not represented in the 3DMM. Similar effects occur when not fine-tuning on NeRSemble, since samples with visible tongue are underrepresented in CelebV-Text.

	Self-reenactment								Cross-reenactment			
	PSNR $\uparrow$	LPIPS $\downarrow$	SSIM $\uparrow$	FID $\downarrow$	CSIM $\uparrow$	AKD $\downarrow$	AED $\downarrow$	APD $\downarrow$	FID $\downarrow$	CSIM $\uparrow$	AED $\downarrow$	APD $\downarrow$
Frozen Denoising UNet	18.11	0.227	0.603	8.38	0.58	0.0072	0.119	0.325	9.59	0.54	0.223	0.494
No Text Control	17.02	0.259	0.566	13.46	0.56	0.0132	0.148	0.380	14.68	0.55	0.249	0.530
No Finetuning on NerSemble	18.72	0.210	0.622	8.15	0.61	0.0061	0.109	0.310	9.00	0.58	0.221	0.486
<b>Ours</b>	18.63	0.212	0.619	7.57	0.62	0.0067	0.110	0.306	8.48	0.57	0.220	0.489

Table 1. Quantitative ablation study of the design choices of our **2D prior**.

	Self-reenactment								Cross-reenactment			
	PSNR $\uparrow$	LPIPS $\downarrow$	SSIM $\uparrow$	FID $\downarrow$	CSIM $\uparrow$	AKD $\downarrow$	AED $\downarrow$	APD $\downarrow$	FID $\downarrow$	CSIM $\uparrow$	AED $\downarrow$	APD $\downarrow$
Ours, $\text{cfg}=1.0$	18.49	0.216	0.611	9.95	0.618	0.00667	0.109	0.310	11.03	0.567	0.2203	0.490
<b>Ours, <math>\text{cfg}=3.0</math></b>	18.63	0.212	0.619	7.57	0.616	0.00669	0.110	0.306	8.48	0.566	0.2201	0.489
Ours, $\text{cfg}=6.0$	18.40	0.221	0.618	8.12	0.594	0.00690	0.116	0.318	9.05	0.548	0.2224	0.491
Ours, $\text{cfg}=10.0$	18.10	0.234	0.611	10.58	0.572	0.00712	0.122	0.329	11.70	0.528	0.2245	0.493

Table 2. Quantitative ablation study of the impact of classifier-free guidance scale (cfg) on our **2D prior**.

leverage text for avatar control.

We found that fine-tuning our model on a mixture of NeRSemble and CelebV-Text after pretraining on CelebV-Text greatly helps in synthesizing tongue articulations (see last column of Figure 3) since these samples are underrepresented in CelebV-Text. However, the quantitative scores slightly deteriorate. We attribute this to a slight overfitting effect on the lighting situation of NeRSemble which causes predictions on samples with uniform backgrounds to have a bias toward this particular lighting setting.

We also evaluate the impact of the classifier-free guidance scale (cfg) on our 2D prior in Figure 3 and Table 2. We found that too small values reduce the faithfulness of extreme expressions while too high values cause oversaturation artifacts. We found that  $\text{cfg}=3$  is a good compromise and also achieves the best FID, PSNR, LPIPS, and SSIM scores in the quantitative self-reenactment evaluation.

### 3.2. Collapse of Dynamic-Target Distillation Approaches for Small Noise Levels

In the main paper, we found that our distillation approach yields better reconstruction results than methods like ImageDream [22], which update the target images at each NeRF optimization step ("*dynamic target*"). We argue that this is because such approaches sample the noise levels

randomly from a specified range, even at the last step of distillation. However, the predictions of the 2D prior at high noise levels typically lack details and exhibit artifacts, particularly for high  $\text{cfg}$  values. Their contribution to the optimization objective bottlenecks the quality of the distillation result. The natural question is if the negative impact of high noise level sampling can be avoided by annealing the upper bound of the sampled noise levels to zero (note that ImageDream caps it to at least 0.5 by default). The result of this experiment is demonstrated in Figure 4. We found that annealing the upper bound of the noise levels to zero makes the distillation diverge. The reason for this is that when performing score distillation sampling (SDS) on small noise levels only, supervision for the low-frequency features like the general shape and outline of the distilled scene is lacking because, at these low-noise levels, only high-frequency details are added by the diffusion prior while the rest is copied over from the input images. However, minor inaccuracies in this process cause the coarse geometry of the 3D reconstruction to drift during the repeated SDS updates while the diffusion prior does not provide correcting gradient directions. As a result, the 3D reconstruction diverges. Only by also sampling high noise levels even at the end of the distillation procedure, guidance on the coarse scene geometry can be achieved, while coming at the cost of reconstruction

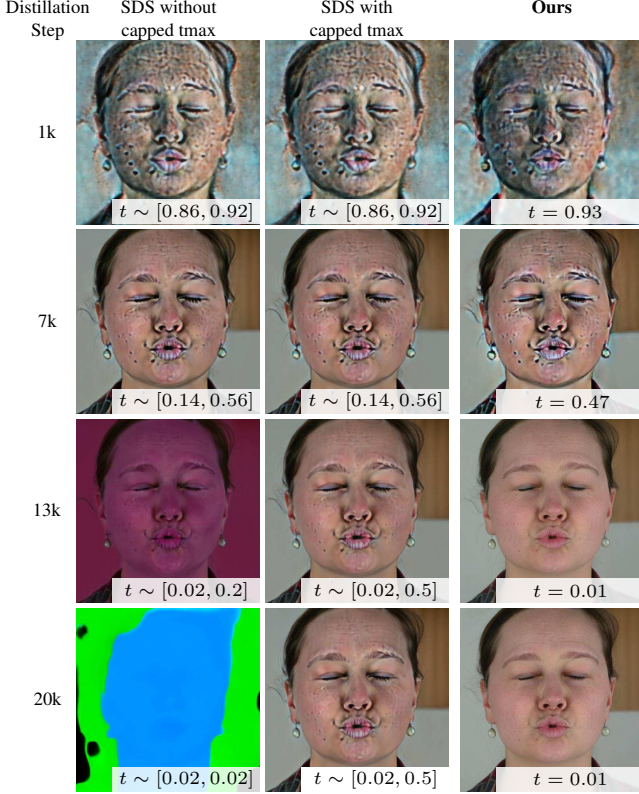


Figure 4. Divergent behavior of score-distillation sampling (SDS) for small noise levels. Typically, SDS-based methods like ImageDream [22] ensure that even towards the end of the distillation procedure high noise levels are sampled, i.e.  $t \sim [0.02, 0.5]$  (see column 2). This bottlenecks the fidelity of the 3D reconstructions and causes artifacts for high cfg values. We found that SDS diverges when not ensuring high-noise levels towards the end of distillation, i.e.  $t \sim [0.02, 0.02]$  (“without capped tmax”, column 1). Our 2-staged approach with deterministic noise levels is able to overcome this limitation (3rd column).

fidelity.

### 3.3. Ablations of Our 3D Distillation Procedure

**Classifier-free guidance scale (cfg)** Figure 6 and Table 3 ablate the impact of the cfg value during distillation. For the quantitative evaluation in Table 3, we follow the same procedure as in the main paper. We find that too small cfg values ( $\sim 5$ ) produce blurry results while too high values ( $\sim 30$ ) result in oversaturation. We chose  $\text{cfg}=19.0$  and found that it yields plausible results of high quality without oversaturation effects.

**Ratio between Stage 1 & 2** Table 4 provides a quantitative ablation study of the impact of the ratios between Stage 1 and Stage 2 during distillation. Please refer to the main paper for a qualitative comparison. We find that increasing the ratio of Stage 2 optimization improves high-

	LPIPS ↓	SSIM ↑	PSNR ↑	MSE ↓
Ours, $\text{cfg}=5.0$	0.199	0.84	22.00	0.0073
Ours, $\text{cfg}=10.0$	0.193	0.83	21.77	0.0076
<b>Ours, <math>\text{cfg}=19.0</math></b>	0.191	0.82	21.53	0.0080
Ours, $\text{cfg}=30.0$	0.191	0.81	21.60	0.0079

Table 3. Quantitative ablation study of the impact of classifier-free guidance scale (cfg) on our **3D distillation** procedure.

Stage 1 / Stage 2	LPIPS ↓	SSIM ↑	PSNR ↑	MSE ↓
100% / 0%	0.27	0.83	20.1	0.012
80% / 20%	0.21	0.82	20.6	0.010
<b>60% / 40%</b>	0.19	0.82	21.5	0.008
30% / 70%	0.18	0.81	22.0	0.007
0% / 100%	0.18	0.81	22.0	0.007

Table 4. Quantitative ablation study of the ratios of Stage 1 and Stage 2 during our **3D distillation**. We use the ratio 60%/40% as the default for our method. While higher ratios of Stage 2 yield better LPIPS, we qualitatively found that it comes at the cost of less consistent reconstructions with semi-transparent artifacts (see main paper).

frequency detail, the LPIPS score improves, yet comes at the cost of reduced consistency and semi-transparent artifacts, the structural similarity index measure (SSIM) worsens. We chose the ratio Stage 1 / Stage 2 of 60%/40% as our default which we found to be a good trade-off between high-frequency details and consistency.

### 3.4. Qualitative Geometry Evaluation

Figure 5 qualitatively visualizes the depth maps of our 3D reconstructions. We observe that our distillation procedure yields plausible geometries with a distinct spatial separation of regions like nose, tongue, mouth cavities, and glasses.

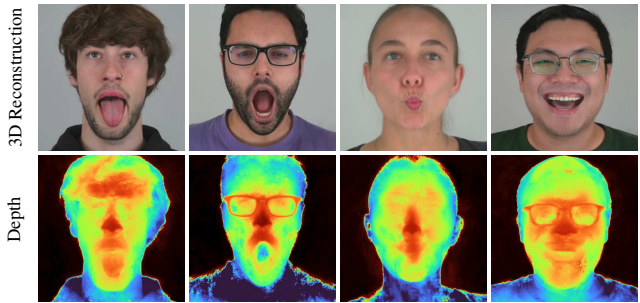


Figure 5. Reconstructed Geometry.

### 3.5. More Qualitative Comparisons of Our 2D Prior

We provide additional qualitative comparisons of our 2D prior with all considered baselines in Figure 7 for self-reenactment and in Figure 8 for cross-reenactment. As observed in the main paper, our 2D prior consistently outper-

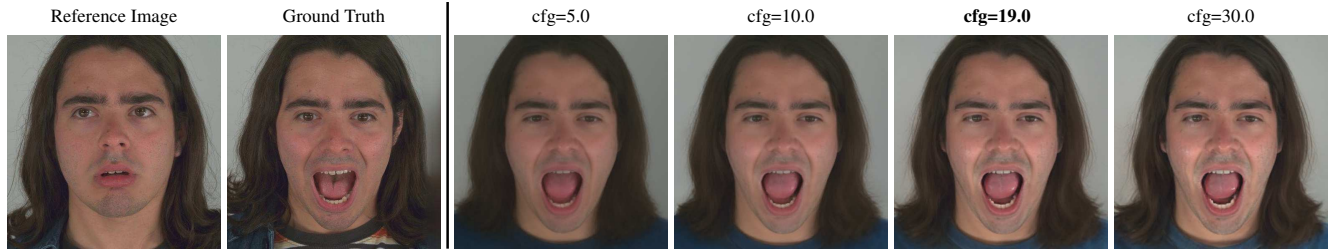


Figure 6. Qualitative ablation study of classifier-free guidance scale (cfg) for our **3D distillation** procedure. Too small values produce blurry results, while too high values cause unnatural oversaturation. We found  $\text{cfg}=19.0$  to be a good compromise and set it as the default for our method.

forms all baselines. It is remarkably robust w.r.t. extreme expressions and poses in the reference and the driving images and produces results with high identity alignment and synthesis quality even on very challenging samples.

#### 4. Ethical Considerations

Our method creates a photo-realistic 3D head reconstruction from a single reference image while providing control over the target pose and expression. It is intended to advance 3D content generation for applications in telecommunications, movie production, and entertainment. Nevertheless, similar to previous work [10, 11, 15, 19, 20, 23], potential misuse in the form of deepfakes is possible. Developing strategies to detect such deepfakes is therefore of critical importance. The field of passive forgery detection enables the identification of deepfakes without explicit watermarking [1, 4, 5, 16, 17]. However, generalized methods [1, 4, 5] have problems in reliably detecting fakes, and therefore cryptographical methods must be used in the future to verify the video’s authenticity.

#### 5. Acknowledgements

This project has received funding from the Max Planck ETH Center for Learning Systems (CLS). Egor Zakharov was funded by the “AI-PERCEIVE” 2021 ERC Consolidator Grant. Further, we would like to thank Phong Tran and Balamurugan Thambiraja for their valuable feedback.

#### References

- [1] Shruti Agarwal, Hany Farid, Tarek El-Gaaly, and Ser-Nam Lim. Detecting deep-fake videos from appearance and behavior. In *2020 IEEE international workshop on information forensics and security (WIFS)*, pages 1–6. IEEE, 2020. 5
- [2] Eric R. Chan, Connor Z. Lin, Matthew A. Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas Guibas, Jonathan Tremblay, Sameh Khamis, Tero Karras, and Gordon Wetzstein. Efficient geometry-aware 3D generative adversarial networks. In *arXiv*, 2021. 2
- [3] Ken Chen, Sachith Seneviratne, Wei Wang, Dongting Hu, Sanjay Saha, Md. Tarek Hasan, Sanka Rasnayaka, Tamasha Malepathirana, Mingming Gong, and Saman Halgamuge. Anifacediff: High-fidelity face reenactment via facial parametric conditioned diffusion models, 2024. 2, 6, 7
- [4] Davide Cozzolino, Justus Thies, Andreas Rössler, Christian Riess, Matthias Nießner, and Luisa Verdoliva. Forensicttransfer: Weakly-supervised domain adaptation for forgery detection. *arXiv preprint arXiv:1812.02510*, 2018. 5
- [5] Davide Cozzolino, Andreas Rössler, Justus Thies, Matthias Nießner, and Luisa Verdoliva. Id-reveal: Identity-aware deepfake video detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 15108–15117, 2021. 5
- [6] Radek Daněček, Michael J Black, and Timo Bolkart. Emoca: Emotion driven monocular face capture and animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20311–20322, 2022. 1
- [7] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4690–4699, 2019. 1
- [8] Yu Deng, Jiaolong Yang, Sicheng Xu, Dong Chen, Yunde Jia, and Xin Tong. Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set. In *IEEE Computer Vision and Pattern Recognition Workshops*, 2019. 1
- [9] Yu Deng, Duomin Wang, and Baoyuan Wang. Portrait4d-v2: Pseudo multi-view data creates better 4d head synthesizer. *arXiv preprint arXiv:2403.13570*, 2024. 2, 6, 7
- [10] Nikita Drobyshev, Jenya Chelisev, Taras Khakhulin, Aleksei Ivakhnenko, Victor Lempitsky, and Egor Zakharov. Megaportraits: One-shot megapixel neural head avatars. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 2663–2671, 2022. 5
- [11] Taras Khakhulin, Vanessa Sklyarova, Victor Lempitsky, and Egor Zakharov. Realistic one-shot mesh-based head avatars. In *European Conference of Computer vision (ECCV)*, pages 345–362. Springer, 2022. 5
- [12] Tobias Kirschstein, Shenhan Qian, Simon Giebenhain, Tim Walter, and Matthias Nießner. Nersemble: Multi-view radiance field reconstruction of human heads. *ACM Trans. Graph.*, 42(4), 2023. 1
- [13] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi.



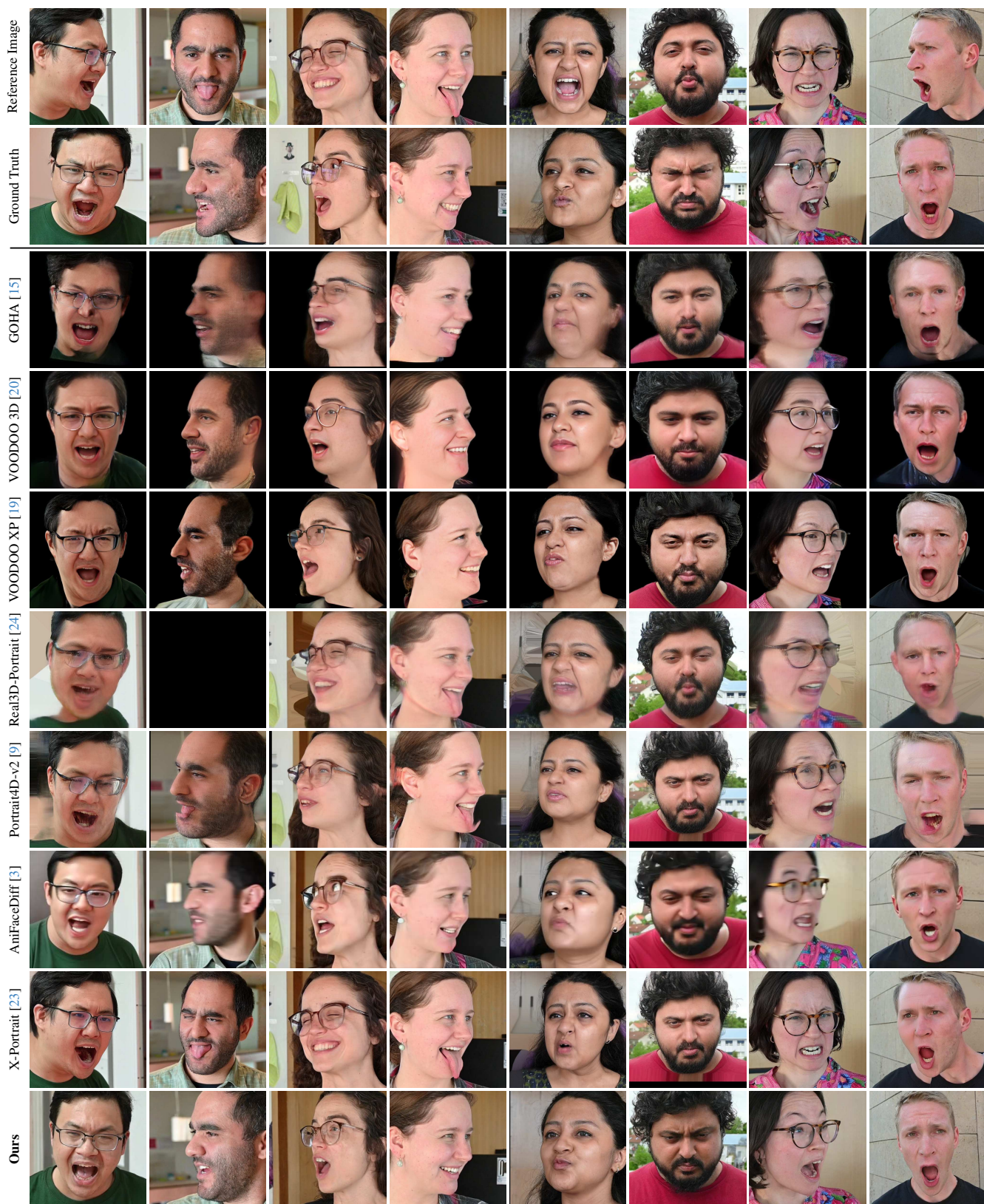


Figure 7. Further qualitative comparisons of our **2D prior** in the self-reenactment scenario. For one sample, Real3D-Portrait’s pose estimator failed, it is marked as a black tile.



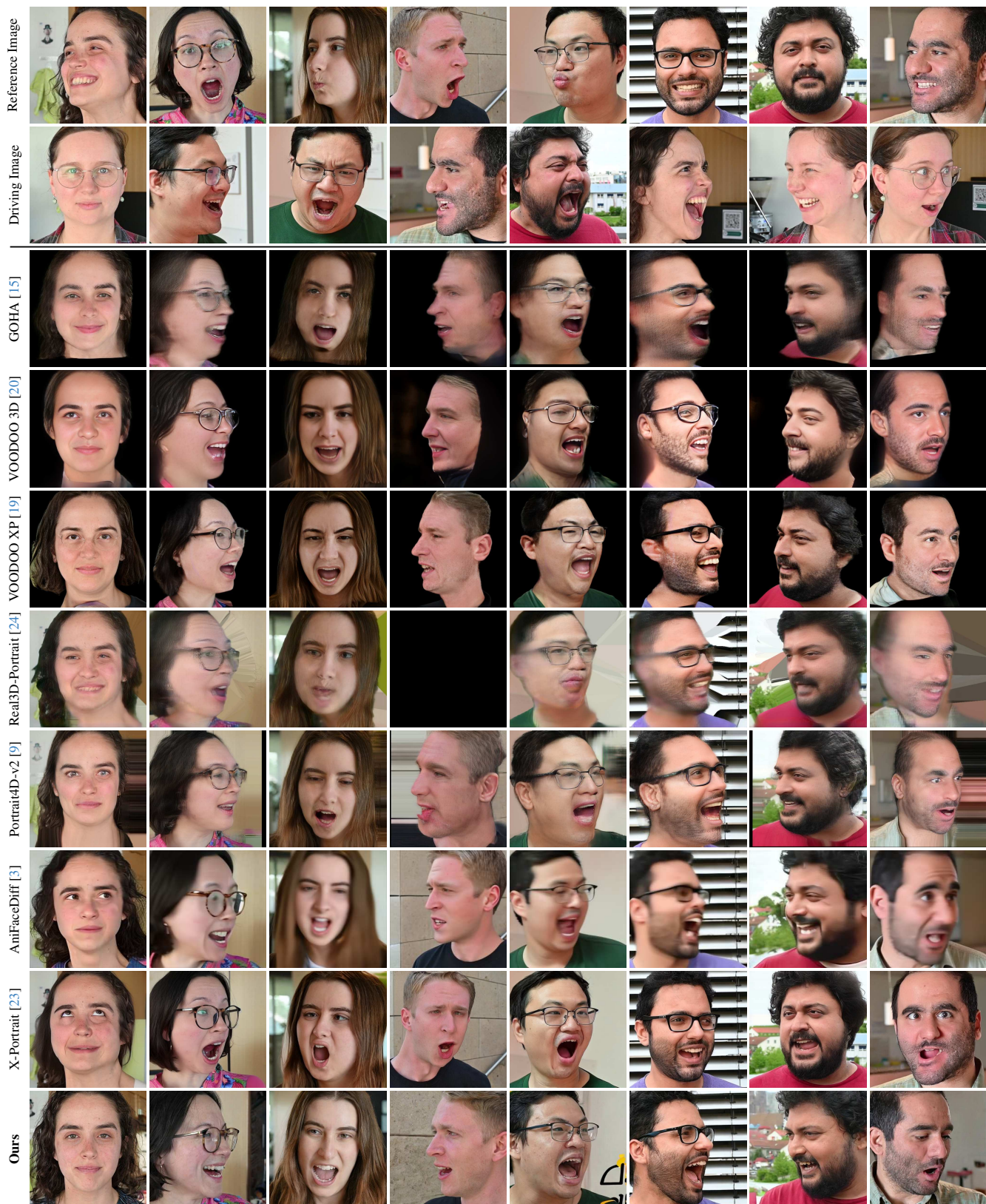


Figure 8. Further qualitative comparisons of our **2D prior** in the cross-reenactment scenario. For one sample, Real3D-Portrait’s pose estimator failed, it is marked as a black tile.

- Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, 2023. [1](#)
- [14] 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. [2](#)
  - [15] Xueting Li, Shalini De Mello, Sifei Liu, Koki Nagano, Umar Iqbal, and Jan Kautz. Generalizable one-shot neural head avatar. *NeurIPS*, 2023. [2](#), [5](#), [6](#), [7](#)
  - [16] Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics: A large-scale video dataset for forgery detection in human faces. *arXiv preprint arXiv:1803.09179*, 2018. [5](#)
  - [17] Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1–11, 2019. [5](#)
  - [18] 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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. [1](#)
  - [19] Phong Tran, Egor Zakharov, Long-Nhat Ho, Liwen Hu, Adilbek Karmanov, Aviral Agarwal, McLean Goldwhite, Ariana Bermudez Venegas, Anh Tuan Tran, and Hao Li. Voodoo xp: Expressive one-shot head reenactment for vr telepresence, 2024. [2](#), [5](#), [6](#), [7](#)
  - [20] Phong Tran, Egor Zakharov, Long-Nhat Ho, Anh Tuan Tran, Liwen Hu, and Hao Li. Voodoo 3d: Volumetric portrait disentanglement for one-shot 3d head reenactment. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. [2](#), [5](#), [6](#), [7](#)
  - [21] Alex Trevithick, Matthew Chan, Michael Stengel, Eric Chan, Chao Liu, Zhiding Yu, Sameh Khamis, Manmohan Chandraker, Ravi Ramamoorthi, and Koki Nagano. Real-time radiance fields for single-image portrait view synthesis. *ACM Transactions on Graphics (TOG)*, 42(4):1–15, 2023. [2](#)
  - [22] Peng Wang and Yichun Shi. Imagedream: Image-prompt multi-view diffusion for 3d generation. *arXiv preprint arXiv:2312.02201*, 2023. [3](#), [4](#)
  - [23] You Xie, Hongyi Xu, Guoxian Song, Chao Wang, Yichun Shi, and Linjie Luo. X-portrait: Expressive portrait animation with hierarchical motion attention, 2024. [2](#), [5](#), [6](#), [7](#)
  - [24] Zhenhui Ye, Tianyun Zhong, Yi Ren, Jiaqi Yang, Weichuang Li, Jiangwei Huang, Ziyue Jiang, Jinzheng He, Rongjie Huang, Jinglin Liu, Chen Zhang, Xiang Yin, Zejun Ma, and Zhou Zhao. Real3d-portrait: One-shot realistic 3d talking portrait synthesis. 2024. [2](#), [6](#), [7](#)
  - [25] Jianhui Yu, Hao Zhu, Liming Jiang, Chen Change Loy, Weidong Cai, and Wayne Wu. CelebV-Text: A large-scale facial text-video dataset. In *CVPR*, 2023. [1](#)