Joker: Conditional 3D Head Synthesis with Extreme Facial Expressions

Malte Prinzler[1,2] Egor Zakharov[3] Vanessa Sklyarova[1,2] Berna Kabadayi[1] Justus Thies[1,4]

[1]Max Planck Institute for Intelligent Systems, Tübingen
[2]Max Planck ETH Center for Intelligent Systems
[3]ETH Zürich [4]Technical University of Darmstadt

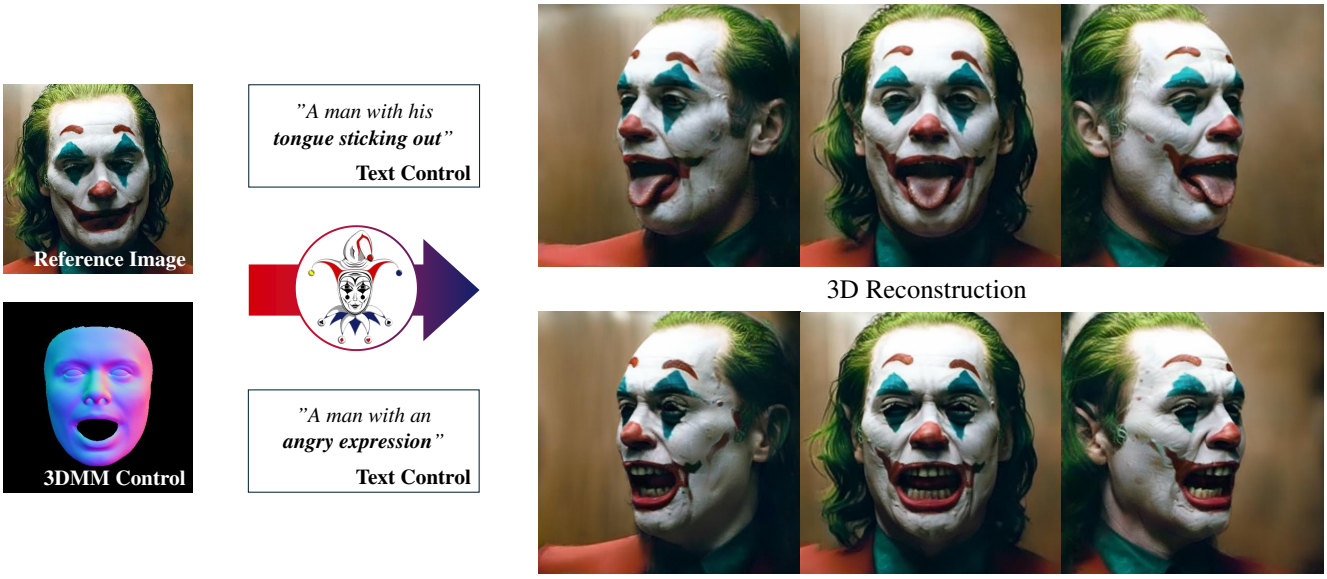

Figure 1. Given a reference image and a target expression defined by a 3DMM and text, Joker generates a 3D model of the subject.

Abstract

We introduce Joker, *a new method for the conditional synthesis of 3D human heads with extreme expressions. Given a single reference image of a person, we synthesize a volumetric human head with the reference's identity and a new expression. We offer control over the expression via a 3D morphable model (3DMM) and textual inputs. This multi-modal conditioning signal is essential since 3DMMs alone fail to define subtle emotional changes and extreme expressions, including those involving the mouth cavity and tongue articulation. Our method is built upon a 2D diffusion-based prior that generalizes well to out-of-domain samples, such as sculptures, heavy makeup, and paintings while achieving high levels of expressiveness. To improve view consistency, we propose a new 3D distillation technique that converts predictions of our 2D prior into a neural radiance field (NeRF). Both the 2D prior and our distillation technique produce state-of-the-art results, which are confirmed by our extensive evaluations. Also, to the best of our knowledge, our method is the first to achieve view-consistent extreme tongue articulation. Project Page*

1. Introduction

Human head avatars have manifold applications in areas such as AR/VR telepresence [39, 48, 62], video games [59, 82], and visual effects [43, 50]. To facilitate downstream applications, techniques for creating 2D head avatars from a single image, often controlled by keypoints, parametric models, and other driving modalities, have been widely studied [19, 20, 30, 57, 66, 69, 72, 73, 75]. However, in many cases, a view-consistent 3D model of a human subject is required to enable rendering in fully virtual environments.

To address this problem, multiple techniques for reconstructing a 3D model of the human subject have been developed [16, 17, 36, 42, 62–64]. Human expressions and appearances, however, have a long-tailed distribution. Mouth cavities and tongues are examples of the areas that are traditionally difficult to capture and, thus, present a challenge for modern avatar systems. Large head rotations are further challenging scenarios, as such examples are missing in most of the existing human-centric datasets [10, 34, 74, 81]. Some approaches [1, 51, 77] addressed this problem by training avatar reconstruction meth-

ods using synthetic datasets of digital human assets rendered via classical graphics pipelines. However, their photorealism and expressiveness still remain subpar. Another group of works [22, 23, 46, 48] extended linear parametric head models [3, 32, 40] with non-linear neural components trained from multi-view datasets [37, 71]. However, collecting such data is expensive, and these approaches can not create avatars with controllable tongues or from a single image. To address these challenges, we propose a novel approach with multi-modal control for extreme expression synthesis. Our method follows an existing line of works on human-centric image synthesis [25, 26, 72] and fine-tunes a pre-trained Stable Diffusion [53] model paired with a ControlNet [78]. We then introduce multi-modal driving inputs to achieve robust and realistic novel-view synthesis of rare and extreme expressions. Our conditioning signal combines the parameters of a 3DMM with a textual prompt. We found that text prompts greatly supplement the control with 3DMM parameters by resolving ambiguities w.r.t. subtle emotional changes and tongue articulations.

View consistency is achieved by optimizing a 3D NeRF [49] from the predictions of our diffusion prior using a novel distillation procedure. Existing 3D distillation approaches [21, 52, 56, 65, 68, 70, 80] exploit 2D diffusion models to predict pseudo-ground-truth target images from noised renders of the current 3D representation. Most of these methods update these target images for every optimization step of the 3D representation – i.e., utilize *dynamic targets*. In a recent concurrent work [21], it was shown that improved performance can be achieved by generating all pseudo ground truth images only once and then optimizing the 3D representation against the generated *static targets*.

We found that neither dynamic nor static target-based approaches yield optimal results for novel-expression 3D distillation. Instead, we propose a new distillation procedure based on *progressively updated targets*. For each time step of a standard DDIM [60] denoising schedule, we use a diffusion-based prior to predict all target images from noised renderings of the 3D representation. We then optimize the 3D representation for several iterations against these target images. Notably, we found it highly beneficial to transition from the dynamically updated target images to static target images at some point during the optimization process. Thus, our proposed progressive distillation method consists of two stages: i) optimization based on dynamically updated targets and ii) optimization based on static targets. We demonstrate that this procedure converges more stably than dynamic-target approaches and is more robust to multi-view inconsistencies than static-target approaches. Ultimately, this yields 3D reconstructions of extreme expressions with high visual fidelity.

To evaluate our method, we collected new benchmark samples that contain extreme expressions in both studio-capture and in-the-wild environments. Thus, we evaluate our method on three datasets: our proposed extreme expression benchmark, CelebV-Text [74], and NeRSemble [37]. We demonstrate an improved performance compared to existing baselines across all these benchmarks. To train our method, we have also collected new metadata for the above-mentioned datasets, such as textual descriptions of the facial expressions and 3DMM fittings.

In summary, we contribute:

- a 2D diffusion model for single-shot extreme expression synthesis with control through text prompts and 3DMM parameters,
- a 3D distillation approach exploiting progressively updated optimization targets to generate photorealistic 3D reconstructions with extreme expressions,
- a new benchmark and metadata for existing training datasets tailored for extreme expression synthesis.

We plan to make our full codebase, validation dataset, and training dataset metadata publicly available.

2. Related work

We aim for controllable 3D head synthesis, given a reference image and a target expression. This target expression is defined in terms of parameters of a 3DMM [32, 40] with additional textual control to allow the synthesis of expressions and appearances that fall outside the space of the 3DMM. Our approach is based on a learned conditional 2D prior which is used for consistent distillation of the 3D head model.

Conditional 2D Head Synthesis. Modern 2D head synthesis methods rely on human-centric image- and video-based datasets to train generative models capable of directly synthesizing images with novel expressions from a single portrait image. Most of these methods utilize GANs [24] or LDMs [53] to enable high realism of the predictions and generalization capabilities of the trained models. GAN-based reenactment methods [5, 18–20, 57, 58, 66, 75, 76] directly predict an output image given a driving signal and a source image. While achieving high visual quality and expressiveness for frontal-facing images, these methods typically suffer from mode collapse, resulting in low generalization capabilities for extreme expressions and head poses.

Diffusion-based approaches [8, 72] resolve some of these limitations by relying on models that were pre-trained on large-scale data, such as Stable Diffusion [53]. These models can be adapted for human reenactment using a separately trained control network [78]. Such an approach achieves a substantially higher degree of generalization than the GAN-based methods trained from scratch using human-centric image and video data. However, its limitations include a substantial change in the identity of the outputs and a lack of explicit viewpoint control [8, 72]. Moreover,

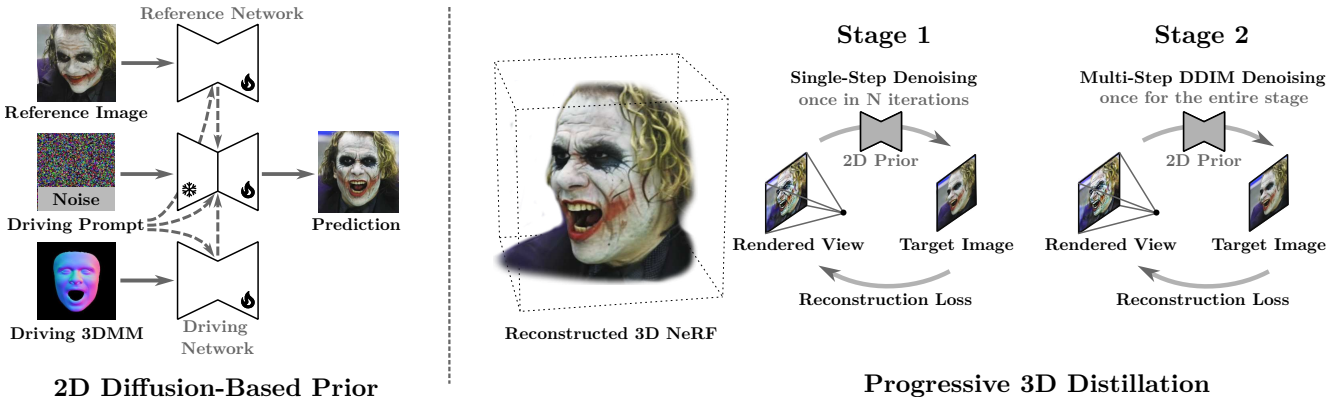

Figure 2. **Method Overview**. We train a 2D diffusion-based prior for novel pose and expression synthesis from a single reference image. It is controlled through text prompts and 3DMM parameters. We leverage this 2D prior to optimize a Neural Radiance Field (NeRF) [49] with a novel two-stage distillation procedure. During Stage 1, the NeRF is optimized against single-step-denoised predictions of the 2D prior that are recalculated every N optimization iterations. In Stage 2, the target images are calculated once in a multi-step denoising process and kept fixed during the NeRF optimization.

these methods use driving modalities such as keypoints [8] or reenacted images produced by a pre-trained GAN-based network [72], which have a limited expressiveness.

Conditional 3D Head Synthesis. 3D-aware head synthesis methods largely address the challenge of view consistency. These methods can also be trained using GAN-based training procedures. Typically, they combine a reconstructed 3D human head model with neural rendering to introduce a high degree of view-consistency [16, 17, 36, 42, 62, 63]. However, these methods have substantial limitations that include a lack of expressiveness and low quality of rendered images. Moreover, they still lack view consistency, especially in high-frequency features, since only the coarse head shape is reconstructed explicitly while super-resolution modules hallucinate the remaining details. These problems are addressed by a growing group of methods that modify pre-trained diffusion models to produce view-consistent renders via viewpoint conditioning [21, 45, 70]. They have been further adapted to human avatar synthesis and can be trained to include explicit pose control [4, 9, 29]. However, since the views are still predicted in image space directly, the 3D consistency of these methods is subpar. While some approaches [25, 72] attempted to resolve this issue with multi-view-aware denoising techniques, the improvements still remain limited.

An alternative approach is the distillation of pre-trained diffusion models into 3D representations using score distillation sampling (SDS) [33, 52, 56, 65, 68], which was used in several previous works on human avatar synthesis [31, 44]. However, these methods either fall short in terms of realism [52] or are unstable w.r.t. the choice of the base diffusion model, i.e. its denoising scheme and hyperparameters, such as classifier-free guidance scale [65, 68].

Contrary to the existing approaches for novel expression synthesis, our method utilizes a progressive optimization strategy of the underlying neural radiance field (NeRF) [49]. First, we utilize dynamically updated targets for supervision via single-step denoising, akin to a classical SDS, which results in a blurry yet consistent reconstruction. Once we achieve a coarse reconstruction, we use it to produce fixed optimization targets using multi-step denoising, which helps to complement the missing high-frequency details. Compared to using fixed ground-truth targets throughout the distillation process, as in the concurrent work [21], we achieve a substantially higher degree of view consistency, especially for the mouth cavity and tongue.

3. Method

Our method takes a portrait image as an input and creates a photo-realistic 3D reconstruction with a driving expression specified via parameters of the Basel Face Model [32] and text prompts, see Figure 2. We train a conditional 2D prior that takes an image of a reference person as input and predicts its novel-view renders with novel expressions (Section 3.1). Exploiting this 2D prior and its predictions, we propose a novel 3D distillation pipeline (Section 3.2) to optimize a view-consistent NeRF.

3.1. 2D Prior for Extreme Expression Synthesis

Our prior model is based on a Stable Diffusion [53] backbone. To convert it into a conditional synthesis model, we follow [7, 9, 25, 72] and train a separate *reference network* to input the information from the reference image into the denoising network. Specifically, we share keys and values between the self-attention layers of the reference network and its denoising counterpart [25]. We then train a ControlNet [78], which we refer to as a *driving network*, to condi-

Figure 3. **3DMM- and text-guided 3D reconstruction.** Through text guidance our model resolves ambiguities in the 3DMM control signal, can formulate tongue articulation, and provides fine-grained emotion control. Note that the 3DMM input is kept fixed for both 3D reconstructions of each row and only the text prompt changes.

tion our model on the parameters of a 3DMM. Compared to previous methods [25, 72] and similar to a concurrent work [9], we utilize mesh-based renders of the normal maps to encode the driving head pose and expression. We have observed that this conditioning style helps to achieve higher view consistency of the results even without view-aware noise or multi-view self-attention techniques [25].

In contrast to existing works [4, 9, 72], we also preserve text-based conditioning from the base denoising model and incorporate it into the reference and driving networks. Text conditioning allows our model to synthesize extreme expressions by supplementing 3DMM-based signals with missing cues, such as tongue movements and emotion-related appearance details. Also, contrary to previous diffusion-based reenactment methods [9, 72], we found it beneficial to fine-tune the decoding part of the denoising network to improve identity preservation. To train this model, we implemented an iterative training pipeline that utilizes a subset of the in-the-wild CelebV-Text dataset [74] followed by a short fine-tuning phase on the NeRSemble dataset [37]. We supplemented these datasets with new metadata that includes textual annotations of the expressions and 3DMM fittings (Appendix A).

3.2. 3D Distillation

Following [56, 65], we use a NeRF [49] formulation to represent our 3D reconstruction. During distillation, the NeRF is rendered under several target views and encoded into the latent space of the diffusion model. Then, noise is added to these renders, and the 2D diffusion prior denoises the latents and decodes them into the target images $\hat{x}_0$ that the NeRF is optimized against.

Following previous works [56, 65, 70], we apply the distillation losses directly in the image domain rather than the latent space. The optimization objective follows [70] and consists of a combination of an L1 distance and a perceptual loss $\mathcal{L}_p$ [79] between the rendered images x and targets $\hat{x}_0$:

$$\mathcal{L}_{\text{recon}} = \mathbb{E}_{\mathbf{c}}\big[\|x - \hat{x}_0\|_1 + \mathcal{L}_p(x, \hat{x}_0)\big], \tag{1}$$

where the expectation is taken over the viewing angles.

Our distillation procedure consists of two stages that differ in the way that the target images are updated. For Stage 1, the target images are repeatedly updated based on the NeRF renderings to improve their multi-view consistency. For Stage 2, the target images are predicted only once and the NeRF is optimized against them until convergence.

Stage 1: Dynamic Target Optimization. During the first stage, we frequently update the target images using the noised NeRF renderings as input to the distillation prior.

We render the NeRF under several target views, apply noise to the renders, and then use the 2D prior to generate the new target images in one denoising step. The NeRF is optimized against the target views for N iterations, after

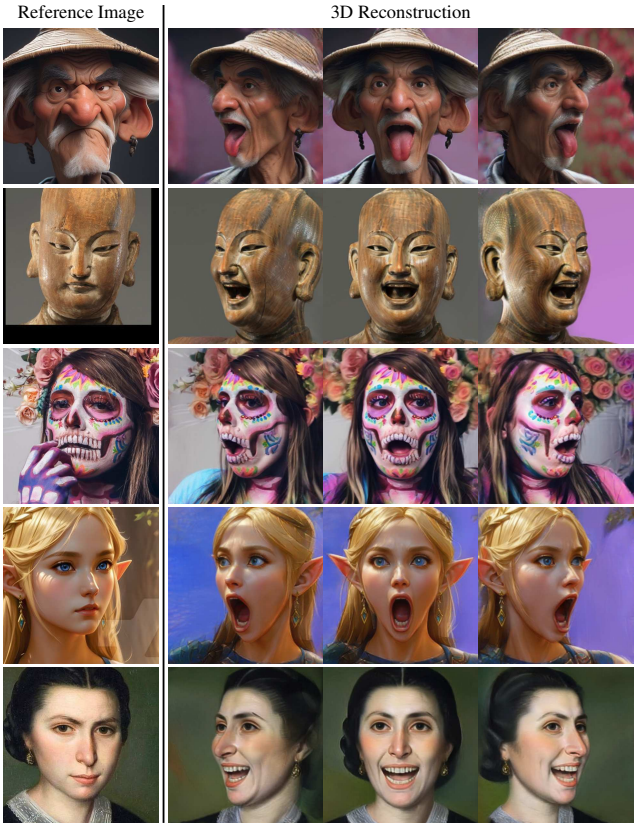

Figure 4. Out-of-distribution 3D reconstruction examples.

which we recalculate the targets and repeat the procedure. We use 3D-consistent NeRF renderings to update the target images to improve their view consistency. The main difference compared to the previous approaches [56, 65, 70] is that we sample the noise levels deterministically following a standard DDIM denoising scheduler with 100 steps, instead of sampling them randomly. Furthermore, the NeRF is optimized against the target images for N iterations before updating them whereas most existing methods update the targets at every NeRF update. We found that this substantially improves the visual quality and consistency of the reconstruction results (see Figure 6).

Stage 2: Fixed Target Optimization. Performing Stage 1 distillation alone tends to drift towards blurry results. The reason for this lies in the NeRF optimization against the target images. During optimization, the NeRF effectively averages over the inconsistencies in the target images and introduces a low-frequency bias. This has a significant impact on the distillation procedure. After the NeRF optimization, its renderings will be more blurry than the target images. The lack of high-frequency details in the NeRF renderings is picked up by the 2D prior and propagates into the updated target images. Optimizing the NeRF against them leads to even more blurry reconstruction results and introduces a positive feedback loop.

To effectively solve this phenomenon, we interrupt Stage 1 at an intermediate noise level (after 60 out of 100 denoising steps) and generate the final target images through standard DDIM sampling with the 40 remaining steps. Afterward, the NeRF is optimized against the final target images until convergence. While the optimization procedure against the fixed target images has a low-frequency bias as well, we avoid the repeated low-frequency feedback loop of Stage 1 and converge to high-quality results.

Note that Stage 2 of our distillation procedure is similar to static-target approaches, with the difference that we start the denoising process not from white noise but from view-consistent renderings of a well-converged NeRF. This largely improves the view consistency of the final target images, while the multi-step denoising procedure generates high-frequency details. As a consequence, the predictions of the 2D prior combine high view consistency and image quality which results in superior NeRF optimization results (see Figure 6).

3.3. Implementation Details

We train the 2D prior starting from pretrained weights of Stable Diffusion v1.5[1]. We use the AdamW optimizer [47] with lr $= 10^{-5}, \beta_1 = 0.9, \beta_2 = 0.999, \epsilon = 10^{-8}, \lambda = 10^{-2}$. Our model is trained on 8 NVIDIA A100 SXM4-80GB with a per-GPU batch size of 10 for 200,000 iterations on CelebV-Text [74] and 30,000 more iterations on an equal mix of CelebV-Text and NerSemble [37]. Please refer to Appendix A for dataset and preprocessing details.

For the 3D distillation, we use the threestudio framework [27] and largely follow the configuration of ImageDream [65]. To generate the target images, we render BFM [32] normal maps on a regular 20×20 grid of the frontal hemisphere with an azimuth $\in [-22.5, 22.5]$ and elevation $\in [-10, 10]$. The images are generated with a classifier-free guidance scale of 19.0. We optimize the NeRF for $N = 130$ iterations between each target image update. During optimization, we randomly sample 64 patches of size 64×64 and gradually increase the resolution of the target images from 64 to 512. The optimization takes approximately 3 hours on a single NVIDIA A100 SXM4-80GB GPU.

4. Experiments

Below we evaluate both our 2D prior and our 3D distillation technique. Please refer to Appendix C for detailed ablation studies and additional experiments. To evaluate the synthesis of the extreme facial expressions, the validation sets of CelebV-Text and NeRSemble alone are not suffi-

[1] huggingface.co/runwayml/stable-diffusion-v1-5

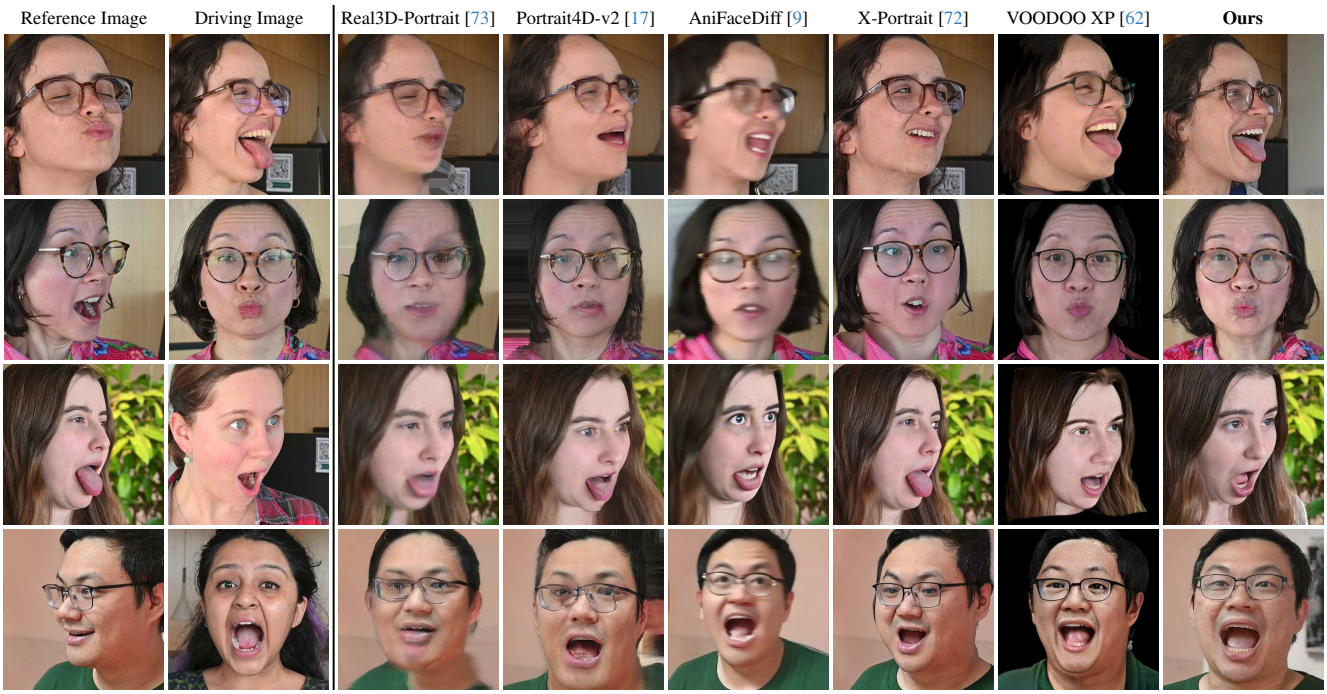

Figure 5. Comparison of our **2D diffusion prior** for self- and cross-reenactment (row 1-2 and 3-4 respectively).

	Self-reenactment								Cross-reenactment			
	PSNR ↑	LPIPS ↓	SSIM ↑	FID ↓	CSIM ↑	AKD ↓	AED ↓	APD ↓	FID ↓	CSIM ↑	AED ↓	APD ↓
GOHA [41]	16.25	0.299	0.572	46.33	0.32	0.015	0.20	0.42	46.82	0.36	0.26	0.51
Real3D-Portrait [73]	16.58	0.322	0.592	31.68	0.46	0.021	0.23	0.47	31.54	0.57	0.30	0.59
Portrait4D-v2 [17]	14.13	0.393	0.494	19.96	0.49	0.081	0.18	0.38	20.71	0.56	0.25	0.52
AniFaceDiff [9]	16.24	0.360	0.577	29.64	0.50	0.034	0.17	0.41	29.34	0.53	0.25	0.52
X-Portrait [72]	14.77	0.357	0.493	10.66	0.61	0.051	0.20	0.68	10.79	0.75*	0.29	0.81
VOODOO 3D [63]	16.11	0.324	0.558	38.63	0.27	0.035	0.19	0.47	39.24	0.30	0.24	0.54
VOODOO XP [62]	13.74	0.397	0.483	24.59	0.49	0.075	0.15	0.45	24.83	0.43	0.21	0.52
Ours	18.63	0.212	0.619	7.57	0.62	0.007	0.11	0.31	8.48	0.57	0.22	0.49

Table 1. Quantitative comparison of our **2D diffusion prior** in self- and cross-reenactment scenarios. *: Note that for extreme pose changes, X-Portrait has a tendency to reproduce the reference image without adopting the driving pose, leading to a high identity similarity score (CSIM) yet poor pose accuracy (APD).

cient: 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 evaluation of extreme expression synthesis, which will be made publicly available to the research community, see Appendix A.

4.1. Evaluation of 2D Prior

We compare our 2D diffusion prior against the following baselines: *GOHA* [41], *VOODOO 3D* [63], *VOODOO XP* [62], *Real3D-Portrait* [73], *Portrait4D-v2* [17], *AniFaceDiff* [9], and *X-Portrait* [72]. We refer to Appendix B for a detailed discussion of those and their implementation details. For the baseline results, we used the official code repository of GOHA, VOODOO 3D, Real3D-Portrait, Portrait4D, and X-Portrait; and obtained the results for AniFaceDiff and VOODOO XP from the authors.

Figure 5 qualitatively compares our 2D diffusion prior with the baselines for the scenario of self- and cross-reenactment. We visualize the best-performing baselines and refer the reader to Figure 14 and Figure 15 for further results. The comparisons are conducted on the newly captured Joker benchmark, which contains in-the-wild scenes with diverse backgrounds, subject ethnicities, and genders. We observe a significant qualitative improvement over all baselines.

These results are confirmed by the quantitative comparison, see Table 1. We evaluate standard metrics such as peak signal-to-noise ratio (PSNR), learned perceptual image patch similarity [79] (LPIPS), structural similarity

	LPIPS ↓	SSIM ↑	PSNR ↑	MSE ↓
ProlificDreamer* [68]	0.43	0.52	15.6	0.028
ImageDream* [65]	0.25	0.79	20.2	0.011
Ours, Stage 1 Only	0.27	0.83	20.1	0.012
Ours, Stage 2 Only	0.18	0.81	22.0	0.007
Ours	0.19	0.82	21.5	0.008

Table 2. Quantitative comparison of our **3D distillation** approach. *: Replacing the method's 2D prior with our model for fair comparison.

index measure [67] (SSIM), and Fréchet inception distance [28] (FID). We measure identity preservation by comparing the cosine similarity between the embeddings of a face recognition network [14] for the predicted and ground truth images (CSIM). Further, we report the average distance between the extracted keypoints (AKD), expression (AED), and pose (APD) parameters using [15]. For the cross-reenactment scenario where no ground truth data is available, we evaluate the CSIM score between the reference image and the prediction and the AED and APD scores between the driving image and prediction.

Since VOODOO 3D, VOODOO XP, and GOHA do not synthesize the background, we mask out the backgrounds of the other methods using MODNet [35] before evaluation. The quantitative comparison is conducted on 10,000 images for self- and cross-reenactment respectively, evenly sampled from the validation sets of CelebV-Text and NeRSemble, our recordings in a studio environment with uniform lighting and background, and our recordings with in-the-wild backgrounds and lighting. Table 1 shows a consistent improvement over all existing methods. Note that VOODOO XP, VOODOO 3D, Real3D-Portrait, GOHA, and Portrait4D are single-step methods that enable real-time inference, while AniFaceDiff, X-Portrait, and our method are diffusion models that require several denoising steps. Further, X-Portrait was trained on video sequences and with a different crop size than our evaluation samples. However, at the time of writing this paper, no code was available to retrain the model.

Note that the very high CSIM identity similarity score of X-Portrait for the cross-reenactment scenario is misleading. We found that for extreme pose changes, X-Portrait tends to reproduce the reference image without adopting the driving pose (see row 3 of Figure 5). Since for the cross-reenactment scenario, we calculate CSIM between the reference image and the prediction, this artifact results in a significantly overestimated CSIM score. This effect is confirmed by the comparatively high pose reconstruction error (APD) of X-Portrait in Table 1. Our method excels in high-fidelity synthesis and identity preservation while being robust w.r.t. extreme expressions and poses both in the reference and driving image.

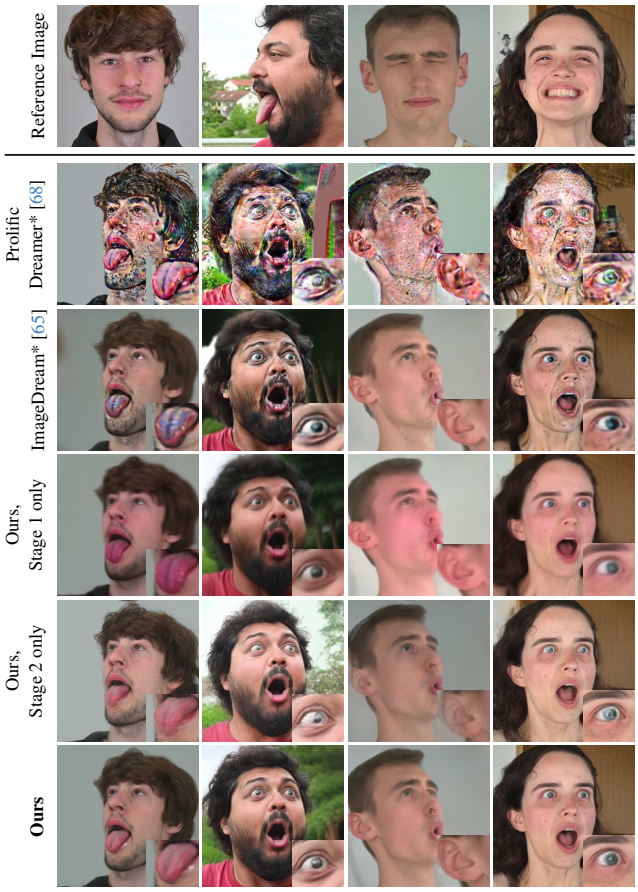

Figure 6. Comparison of **3D reconstructions** from different distillation procedures. *: Replacing the method's 2D prior with our model for fair comparison.

4.2. Evaluation of 3D Distillation

We compare our novel distillation approach against two state-of-the-art baselines: *ImageDream* [65] and *ProlificDreamer* [68]. *ImageDream* [65] uses a diffusion prior that predicts consistent multi-view images given a reference image and exploits this prior to perform multi-view score distillation sampling. *ProlificDreamer* [68] generalizes score distillation sampling to variational score distillation by treating the NeRF renderings as random variables approximated by a pose-conditioned diffusion model that is fine-tuned on the NeRF renderings during distillation. For ImageDream, we use the official code base, for ProlificDreamer we use the threestudio implementation [27]. For a fair comparison, we replace the 2D diffusion priors of both baselines with our own prior. We further compare against versions of our method that only use Stage 1 and Stage 2 respectively. Note that the Stage-2-only setting is similar to Cat3D [21]. However, Cat3D only considers a novel view synthesis scenario where reference images of the same scene are given. Instead, in our scenario the refer-

ence and output images differ drastically due to strong pose and expression changes.

Figure 6 presents a qualitative comparison of the distillation procedures on samples from the NerSemble validation set and our self-captured *Joker benchmark* with in-the-wild scenarios. We observe that ProlificDreamer exhibits instabilities during distillation. They are caused by inaccuracies in the fine-tuned diffusion prior that approximates the probability distribution of the NeRF renderings. ImageDream converges more stably, yet artifacts remain since even at the end of the distillation procedure, denoising timesteps are sampled randomly including high noise levels leading to inaccurate estimates of the target images for the optimization of the NeRF. These artifacts are resolved using the Stage 1 optimization of our approach that follows a deterministic denoising schedule. However, using only Stage 1 yields blurry synthesis results. As discussed in Section 3.2, the low-frequency bias of Stage 1 stems from a repeated interplay between NeRF optimization and target image prediction. When performing Stage 2 only, we suppress this effect and can distill a NeRF with high-frequency details. However, note that in Stage 2 all target images are generated at once and the NeRF is optimized against fixed targets. The inconsistencies in these images cause synthesis artifacts like semi-transparencies and misalignment artifacts in the eyes and around the silhouette. Please refer to the suppl. video for a dynamic comparison of the distillation results.

In Table 2, we also perform a quantitative comparison on 30 samples from the NeRSemble validation set which provides multi-view ground truth images. Our distillation approach consistently improves over the baselines. We observe that while using both stages of our method qualitatively yielded the best combination of view consistency and high-frequency detail, using only Stage 2 even improves the scores slightly, however, at the cost of view consistency. Please refer to the suppl. video for a dynamic visualization of this effect. Appendix C.3 further provides an evaluation of the impact of the classifier-free guidance scale and the ratio between Stage 1 and Stage 2. We observe that our distillation generates plausible geometry and generalizes well to challenging out-of-distribution samples (see Figure 4).

4.3. Text-Guided Expression Synthesis

Figure 3 demonstrates the effectiveness of using text prompts to control the 3D reconstruction and disambiguate the control through 3DMM parameters: Each row presents two reconstruction results that use the same reference image and 3DMM parameters but different text prompts. We find that text prompts provide an intuitive control mechanism to specify the target emotion and tongue articulation. Please refer to Figure 10 and Table 3 for comparisons against a model without text control.

Figure 7. Failure cases of our method.

4.4. Limitations

Figure 7 visualizes failure cases of our method. We observe that implausible colors may be synthesized for challenging out-of-distribution samples in face regions that are not visible in the reference image. In rare cases, we find that even for in-distribution samples the high cfg value (19.0) during distillation causes unnatural colorizations. Please refer to Appendix C.3 for an ablation study on this parameter during distillation. For samples with a uniform background, we observe a tendency of our model to project them to the NerSemble lighting setting: compare rows 2 (NeRSemble) and 3 (Joker benchmark) of Figure 7. Lastly, particularly for dark curly hair, we observe a low-frequency bias. While using only Stage 2 of our distillation procedure can reduce this effect, this comes at the cost of reduced 3D consistency. Further, note that we only consider static scenes in our work. Extending it to 4D avatar synthesis is a fascinating topic for future research.

5. Conclusion

We introduced *Joker*, a novel method for conditional 3D human head synthesis with extreme expressions from a single reference image. Based on control through 3DMM parameters and text prompts, our method produces high-quality results and generalizes well to out-of-distribution samples. The foundation of this approach is a 2D diffusion-based prior which is learned on in-the-wild imagery of human faces. We leverage this prior to progressively distill a 3D volumetric representation of the target subject with a different facial expression. The textual description allows us to specify the facial expression state beyond the parameters of the 3DMM, including subtle emotional changes, as well as extreme expressions with protruding tongue. We believe that *Joker* is a stepping stone for creating high-resolution 3D content of people with a high degree of identity preservation and emotional expressiveness.

6. 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] ItSeez3D AvatarSDK, https://avatarsdk.com. 1

[2] 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. 17

[3] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3d faces. *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, 1999. 2

[4] Stella Bounareli, Christos Tzelepis, Vasileios Argyriou, Ioannis Patras, and Georgios Tzimiropoulos. Diffusionact: Controllable diffusion autoencoder for one-shot face reenactment. 2024. 3, 4

[5] Egor Burkov, Igor Pasechnik, Artur Grigorev, and Victor Lempitsky. Neural head reenactment with latent pose descriptors. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13786–13795, 2020. 2

[6] 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. 14

[7] Di Chang, Yichun Shi, Quankai Gao, Jessica Fu, Hongyi Xu, Guoxian Song, Qing Yan, Xiao Yang, and Mohammad Soleymani. Magicdance: Realistic human dance video generation with motions & facial expressions transfer. *arXiv preprint arXiv:2311.12052*, 2023. 3

[8] Di Chang, Yichun Shi, Quankai Gao, Jessica Fu, Hongyi Xu, Guoxian Song, Qing Yan, Yizhe Zhu, Xiao Yang, and Mohammad Soleymani. Magicpose: Realistic human poses and facial expressions retargeting with identity-aware diffusion, 2024. 2, 3

[9] 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. 3, 4, 6, 14, 18, 19

[10] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. Voxceleb2: Deep speaker recognition. *Interspeech 2018*, 2018. 1

[11] Davide Cozzolino, Justus Thies, Andreas Rössler, Christian Riess, Matthias Nießner, and Luisa Verdoliva. Forensictransfer: Weakly-supervised domain adaptation for forgery detection. *arXiv preprint arXiv:1812.02510*, 2018. 17

[12] 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. 17

[13] 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. 13

[14] 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. 7, 13

[15] 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. 7, 13

[16] Yu Deng, Duomin Wang, Xiaohang Ren, Xingyu Chen, and Baoyuan Wang. Portrait4d: Learning one-shot 4d head avatar synthesis using synthetic data. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 1, 3

[17] 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, 3, 6, 14, 18, 19

[18] Michail Christos Doukas, Stefanos Zafeiriou, and Viktoriia Sharmanska. Headgan: One-shot neural head synthesis and editing. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 2

[19] Nikita Drobyshev, Jenya Chelishev, 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. 1, 17

[20] Nikita Drobyshev, Antoni Bigata Casademunt, Konstantinos Vougioukas, Zoe Landgraf, Stavros Petridis, and Maja Pantic. Emoportraits: Emotion-enhanced multimodal one-shot head avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 1, 2

[21] Ruiqi Gao*, Aleksander Holynski*, Philipp Henzler, Arthur Brussee, Ricardo Martin-Brualla, Pratul P. Srinivasan, Jonathan T. Barron, and Ben Poole*. Cat3d: Create anything in 3d with multi-view diffusion models. *arXiv*, 2024. 2, 3, 7

[22] Simon Giebenhain, Tobias Kirschstein, Markos Georgopoulos, Martin Rünz, Lourdes Agapito, and Matthias Nießner. Mononphm: Dynamic head reconstruction from monocular videos. *ArXiv*, abs/2312.06740, 2023. 2

[23] Simon Giebenhain, Tobias Kirschstein, Markos Georgopoulos, Martin Rünz, Lourdes Agapito, and Matthias Nießner. Learning neural parametric head models. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[24] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 2

[25] Yuming Gu, Hongyi Xu, You Xie, Guoxian Song, Yichun Shi, Di Chang, Jing Yang, and Lingjie Luo. Diffportrait3d: Controllable diffusion for zero-shot portrait view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2, 3, 4

[26] Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *International Conference on Learning Representations*, 2024. 2

[27] Yuan-Chen Guo, Ying-Tian Liu, Ruizhi Shao, Christian Laforte, Vikram Voleti, Guan Luo, Chia-Hao Chen, Zi-Xin Zou, Chen Wang, Yan-Pei Cao, and Song-Hai Zhang. threestudio: A unified framework for 3d content generation. https://github.com/threestudio-project/threestudio, 2023. 5, 7

[28] 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. 7

[29] Hsuan-I Ho, Jie Song, and Otmar Hilliges. Sith: Single-view textured human reconstruction with image-conditioned diffusion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 3

[30] Fa-Ting Hong, Longhao Zhang, Li Shen, and Dan Xu. Depth-aware generative adversarial network for talking head video generation. 2022. 1

[31] Yangyi Huang, Hongwei Yi, Yuliang Xiu, Tingting Liao, Jiaxiang Tang, Deng Cai, and Justus Thies. TeCH: Text-guided Reconstruction of Lifelike Clothed Humans. In *International Conference on 3D Vision (3DV)*, 2024. 3

[32] *A 3D Face Model for Pose and Illumination Invariant Face Recognition*, Genova, Italy, 2009. IEEE. 2, 3, 5

[33] Chenhan Jiang, Yihan Zeng, Tianyang Hu, Songcun Xu, Wei Zhang, Wei Xu, and Dit-Yan Yeung. Jointdreamer: Ensuring geometry consistency and text congruence in text-to-3d generation via joint score distillation. In *ECCV*, 2024. 3

[34] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1

[35] Zhanghan Ke, Jiayu Sun, Kaican Li, Qiong Yan, and Rynson W.H. Lau. Modnet: Real-time trimap-free portrait matting via objective decomposition. In *AAAI*, 2022. 7

[36] 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. 1, 3, 17

[37] 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. 2, 4, 5, 13

[38] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, 2023. 13

[39] Ruilong Li, Kyle Olszewski, Yuliang Xiu, Shunsuke Saito, Zeng Huang, and Hao Li. Volumetric human teleportation. In *ACM SIGGRAPH 2020 Real-Time Live!*, 2020. 1

[40] 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, 14

[41] Xueting Li, Shalini De Mello, Sifei Liu, Koki Nagano, Umar Iqbal, and Jan Kautz. Generalizable one-shot neural head avatar. *NeurIPS*, 2023. 6, 14, 17, 18, 19

[42] Xueting Li, Shalini De Mello, Sifei Liu, Koki Nagano, Umar Iqbal, and Jan Kautz. Generalizable one-shot neural head avatar. *Advances in Neural Information Processing Systems*, 36, 2023. 1, 3

[43] Yixuan Li, Chao Ma, Yichao Yan, Wenhan Zhu, and Xiaokang Yang. 3d-aware face swapping. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12705–12714, 2023. 1

[44] Tingting Liao, Hongwei Yi, Yuliang Xiu, Jiaxiang Tang, Yangyi Huang, Justus Thies, and Michael J Black. Tada! text to animatable digital avatars. *ArXiv*, 2023. 3

[45] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object, 2023. 3

[46] Stephen Lombardi, Jason M. Saragih, Tomas Simon, and Yaser Sheikh. Deep appearance models for face rendering. *ACM Transactions on Graphics (TOG)*, 37:1 – 13, 2018. 2

[47] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2019. 5

[48] Shugao Ma, Tomas Simon, Jason Saragih, Dawei Wang, Yuecheng Li, Fernando De La Torre, and Yaser Sheikh. Pixel codec avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 64–73, 2021. 1, 2

[49] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020. 2, 3, 4

[50] Jacek Naruniec, Leonhard Helminger, Christopher Schroers, and Romann M. Weber. High-resolution neural face swapping for visual effects. *Computer Graphics Forum*, 39, 2020. 1

[51] Pinscreen, 2024. Pinscreen Avatar Neo, https://www.avatarneo.com. 1

[52] Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv*, 2022. 2, 3

[53] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021. 2, 3

[54] 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. 17

[55] Andreas Rossler, 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. 17

[56] Yichun Shi, Peng Wang, Jianglong Ye, Long Mai, Kejie Li, and Xiao Yang. Mvdream: Multi-view diffusion for 3d generation. *arXiv:2308.16512*, 2023. 2, 3, 4, 5

[57] 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. 1, 2

[58] Aliaksandr Siarohin, Oliver Woodford, Jian Ren, Menglei Chai, and Sergey Tulyakov. Motion representations for articulated animation. In *CVPR*, 2021. 2

[59] Vanessa Sklyarova, Egor Zakharov, Otmar Hilliges, Michael J. Black, and Justus Thies. Haar: Text-conditioned generative model of 3d strand-based human hairstyles. *ArXiv*, abs/2312.11666, 2023. 1

[60] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021. 2

[61] 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. 13

[62] 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. 1, 3, 6, 14, 17, 18, 19

[63] 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. 3, 6, 14, 17, 18, 19

[64] 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. 1, 14

[65] Peng Wang and Yichun Shi. Imagedream: Image-prompt multi-view diffusion for 3d generation. *arXiv preprint arXiv:2312.02201*, 2023. 2, 3, 4, 5, 7, 15, 16

[66] Ting-Chun Wang, Arun Mallya, and Ming-Yu Liu. One-shot free-view neural talking-head synthesis for video conferencing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10039–10049, 2021. 1, 2

[67] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4): 600–612, 2004. 7

[68] Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. *arXiv preprint arXiv:2305.16213*, 2023. 2, 3, 7

[69] Olivia Wiles, A. Sophia Koepke, and Andrew Zisserman. X2face: A network for controlling face generation by using images, audio, and pose codes. *ArXiv*, abs/1807.10550, 2018. 1

[70] Rundi Wu, Ben Mildenhall, Philipp Henzler, Keunhong Park, Ruiqi Gao, Daniel Watson, Pratul P. Srinivasan, Dor Verbin, Jonathan T. Barron, Ben Poole, and Aleksander Holynski. Reconfusion: 3d reconstruction with diffusion priors. *arXiv*, 2023. 2, 3, 4, 5

[71] Cheng-hsin Wuu, Ningyuan Zheng, Scott Ardisson, Rohan Bali, Danielle Belko, Eric Brockmeyer, Lucas Evans, Timothy Godisart, Hyowon Ha, Xuhua Huang, Alexander Hypes, Taylor Koska, Steven Krenn, Stephen Lombardi, Xiaomin Luo, Kevyn McPhail, Laura Millerschoen, Michal Perdoch, Mark Pitts, Alexander Richard, Jason Saragih, Junko Saragih, Takaaki Shiratori, Tomas Simon, Matt Stewart, Autumn Trimble, Xinshuo Weng, David Whitewolf, Chenglei Wu, Shoou-I Yu, and Yaser Sheikh. Multiface: A dataset for neural face rendering. In *arXiv*, 2022. 2

[72] You Xie, Hongyi Xu, Guoxian Song, Chao Wang, Yichun Shi, and Linjie Luo. X-portrait: Expressive portrait animation with hierarchical motion attention, 2024. 1, 2, 3, 4, 6, 14, 17, 18, 19

[73] 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. 1, 6, 14, 18, 19

[74] 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, 2, 4, 5, 13

[75] Egor Zakharov, Aliaksandra Shysheya, Egor Burkov, and Victor Lempitsky. Few-shot adversarial learning of realistic neural talking head models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9459–9468, 2019. 1, 2

[76] Egor Zakharov, Aleksei Ivakhnenko, Aliaksandra Shysheya, and Victor Lempitsky. Fast bi-layer neural synthesis of one-shot realistic head avatars. In *European Conference on Computer Vision*, pages 524–540. Springer, 2020. 2

[77] Bowen Zhang, Yiji Cheng, Chunyu Wang, Ting Zhang, Jiaolong Yang, Yansong Tang, Feng Zhao, Dong Chen, and Baining Guo. Rodinhd: High-fidelity 3d avatar generation with diffusion models. *arXiv preprint arXiv:2407.06938*, 2024. 1

[78] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models, 2023. 2, 3

[79] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 4, 6

[80] Yufeng Zheng, Xueting Li, Koki Nagano, Sifei Liu, Otmar Hilliges, and Shalini De Mello. A unified approach for text- and image-guided 4d scene generation. In *CVPR*, 2024. 2

[81] 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 *European conference on computer vision*, pages 650–667. Springer, 2022. 1

[82] Luyang Zhu, Konstantinos Rematas, Brian Curless, Steven M. Seitz, and Ira Kemelmacher-Shlizerman. Reconstructing nba players. In *European Conference on Computer Vision*, 2020. 1

Joker: Conditional 3D Head Synthesis with Extreme Facial Expressions

Supplementary Material

Figure 8. Further 3D reconstructions of our method on a diverse set of examples including out-of-distribution samples.

A. Datasets

As described in the main paper, we use the datasets CelebV-Text [74] and NeRSemble [37] 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 [74] 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 [61] 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[2] [13] 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 [15] and automatically annotated with BFM parameters and text captions using Deep3DFaceRecon [15] and Blip2 [38]. 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 [14] 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 [37] 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.

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 back-

[2] https://github.com/radekd91/inferno/tree/master/inferno_apps/FaceReconstruction

Figure 9. 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.

ground 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 9. 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.

B. Description of the baseline methods

VOODOO 3D [63] finetunes a pretrained model [64] 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* [62] 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* [73] 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* [17] combines a modified EG3D [6] pipeline with a control mechanism through the FLAME 3DMM [40]. *GOHA* [41] uses a 3DMM to control facial expressions by mapping 3DMM parameters to residuals of a tri-plane representation of the face. *AniFaceDiff* [9] follows a similar approach as our method, yet instead of using a ControlNet, they encode normal maps of FLAME [40] 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* [72] 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.

C. Additional Experiments

C.1. Ablation Study of Our 2D Prior

We ablate the design choices of our 2D prior in Table 3 and Figure 10. In contrast to X-Portrait, we unfreeze the upsampling 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 10).

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 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

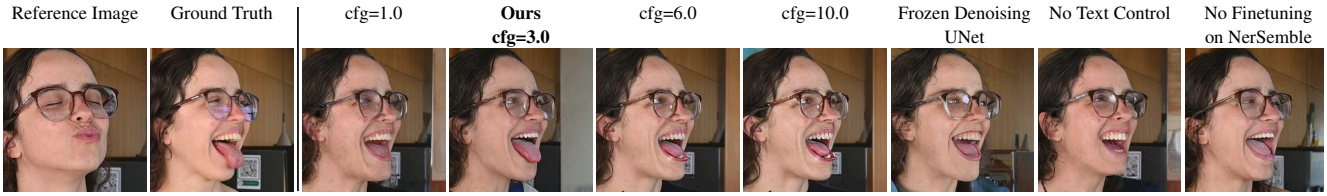

Figure 10. Qualitative ablation study of our **2D prior**. Too small classifier-free guidance scale values (cfg) reduce the faithfulness of extreme expressions, while too high values cause oversaturation and artifacts. We find that cfg=3 yields the best trade-off. Not training the upsampling layers of 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 ↑	LPIPS ↓	SSIM ↑	FID ↓	CSIM ↑	AKD ↓	AED ↓	APD ↓	FID ↓	CSIM ↑	AED ↓	APD ↓
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
Ours	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 3. Quantitative ablation study of the design choices of our **2D prior**.

	Self-reenactment								Cross-reenactment			
	PSNR ↑	LPIPS ↓	SSIM ↑	FID ↓	CSIM ↑	AKD ↓	AED ↓	APD ↓	FID ↓	CSIM ↑	AED ↓	APD ↓
Ours, 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
Ours, cfg=3.0	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, 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, 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 4. Quantitative ablation study of the impact of classifier-free guidance scale (cfg) on our **2D prior**.

last column of Figure 10) 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 10 and Table 4. We found that too small values reduce the faithfulness of extreme expressions while too high values cause oversaturation artifacts. We found that cfg=3 is a good compromise and also achieves the best FID, PSNR, LPIPS, and SSIM scores in the quantitative self-reenactment evaluation.

C.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 [65], 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 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 11. 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 fidelity.

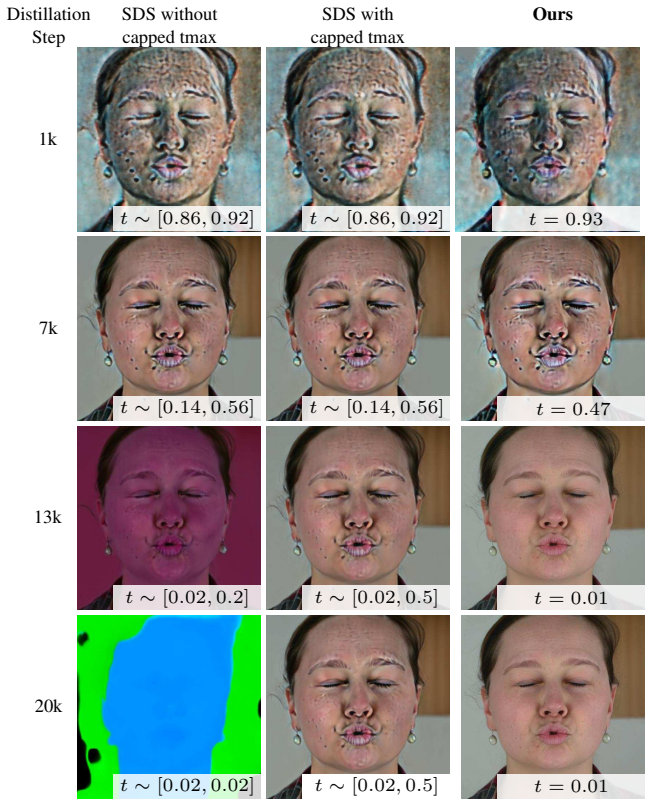

Figure 11. Divergent behavior of score-distillation sampling (SDS) for small noise levels. Typically, SDS-based methods like ImageDream [65] ensure that even towards the end of the distillation procedure high noise levels are sampled, i.e. $t \sim [0.02, 0.5]$ (see column 2). This bottlenecks the fidelity 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).

C.3. Ablations of Our 3D Distillation Procedure

Classifier-free guidance scale (cfg) Figure 13 and Table 5 ablate the impact of the cfg value during distillation. For the quantitative evaluation in Table 5, 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 cfg=19.0 and found that it yields plausible results of high quality without oversaturation effects.

Ratio between Stage 1 & 2 Table 6 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-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.

	LPIPS ↓	SSIM ↑	PSNR ↑	MSE ↓
Ours, cfg=5.0	0.199	0.84	22.00	0.0073
Ours, cfg=10.0	0.193	0.83	21.77	0.0076
Ours, cfg=19.0	0.191	0.82	21.53	0.0080
Ours, cfg=30.0	0.191	0.81	21.60	0.0079

Table 5. 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
60% / 40%	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 6. 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).

C.4. Qualitative Geometry Evaluation

Figure 12 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.

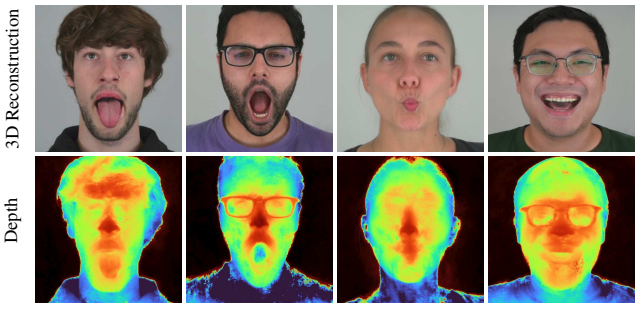

Figure 12. Reconstructed Geometry.

C.5. More Qualitative Comparisons of Our 2D Prior

We provide additional qualitative comparisons of our 2D prior with all considered baselines in Figure 14 for self-reenactment and in Figure 15 for cross-reenactment. As observed in the main paper, our 2D prior consistently outperforms all baselines. It is remarkably robust w.r.t. extreme

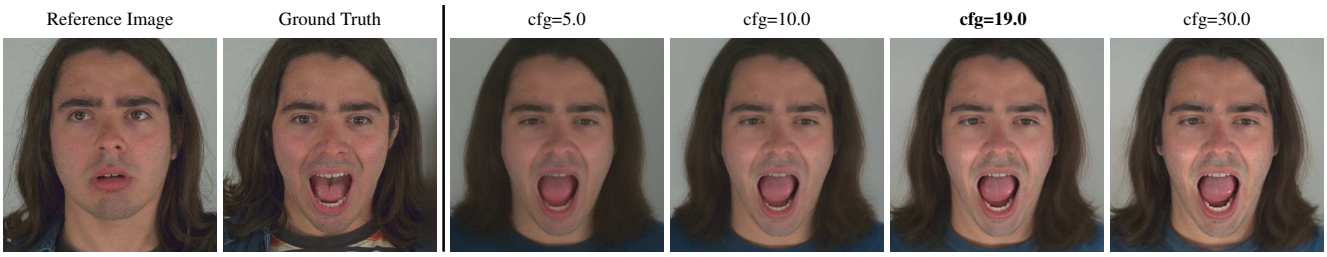

Figure 13. Qualitative ablation study of classifier-free guidance scale (cfg) for our **3D distillation** procedure. Too small values produce blurry results, while too high values cause unnatural oversaturation. We found cfg=19.0 to be a good compromise and set it as the default for our method.

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.

D. 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 [19, 36, 41, 62, 63, 72], 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 [2, 11, 12, 54, 55]. However, generalized methods [2, 11, 12] have problems in reliably detecting fakes, and therefore cryptographical methods must be used in the future to verify the video's authenticity.

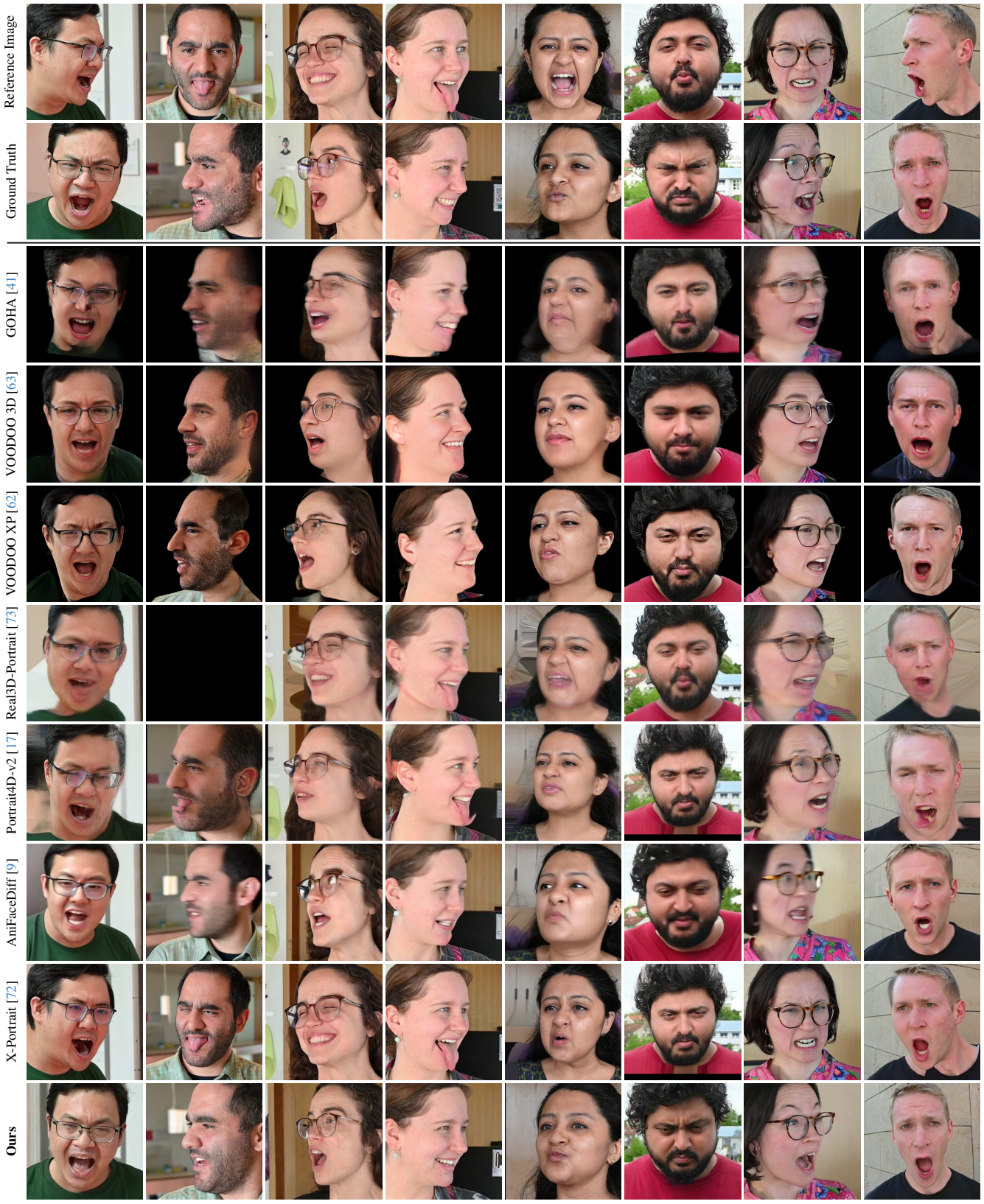

Figure 14. Further qualitative comparisons of our **2D prior** in the self-reenactment scenario. For one sample, Real3D-Portrait's pose estimator failed, it is marked as a black tile.

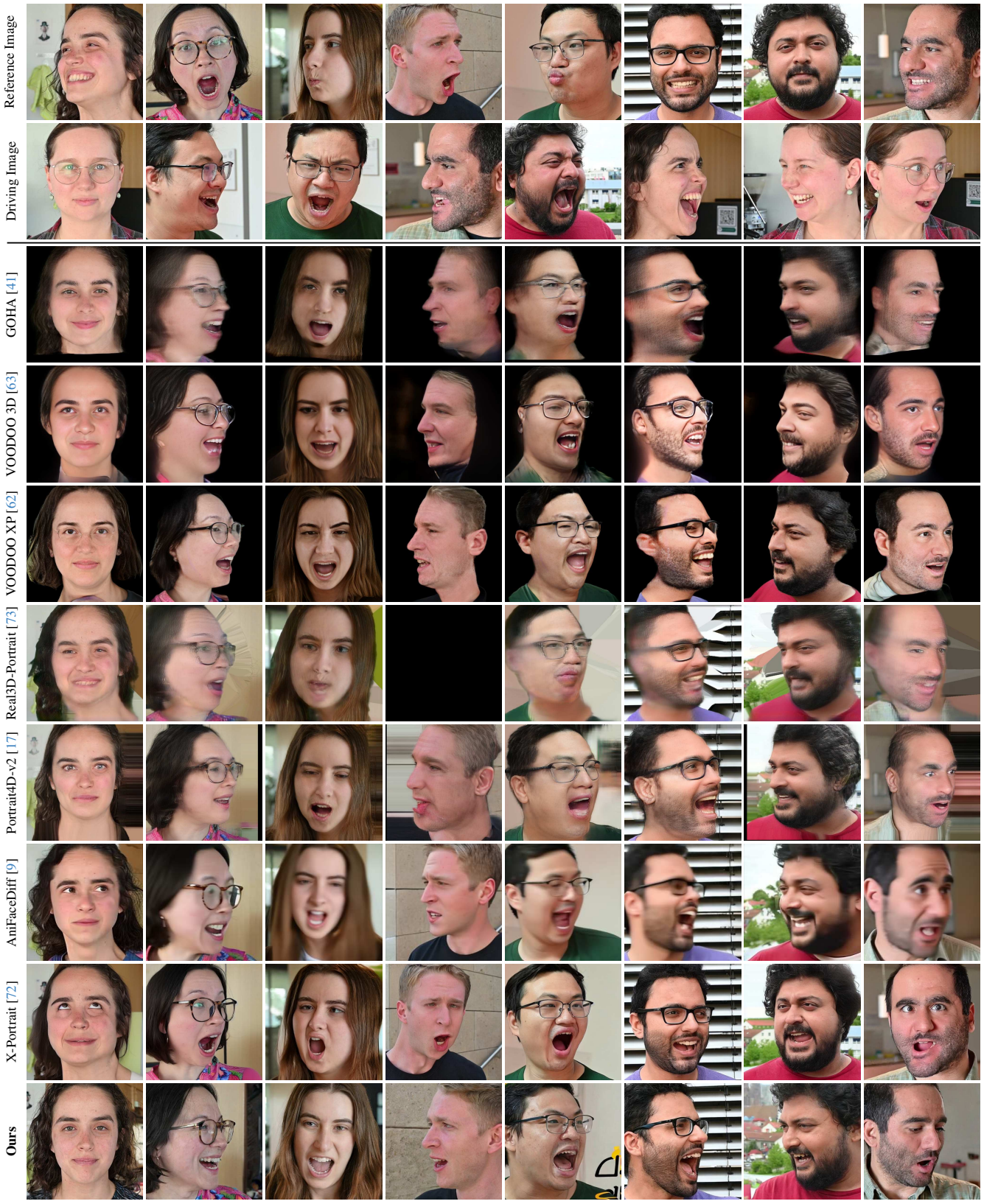

Figure 15. Further qualitative comparisons of our **2D prior** in the cross-reenactment scenario. For one sample, Real3D-Portrait's pose estimator failed, it is marked as a black tile.