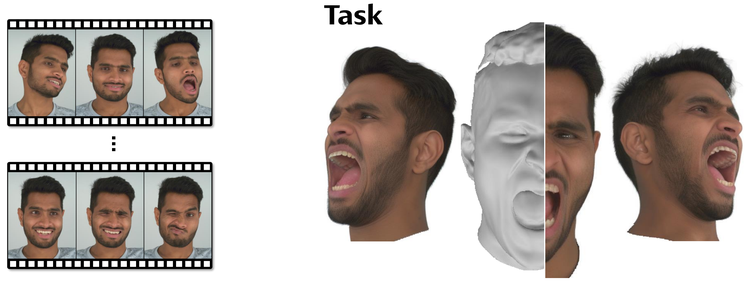
DiffusionAvatars: Deferred Diffusion for High-fidelity 3D Head Avatars

Tobias Kirschstein[1], Simon Giebenhain[1], Matthias Nießner[1]

[1]Technical University of Munich

CVPR SEATTLE, WA JUNE 17-21, 2024

Task



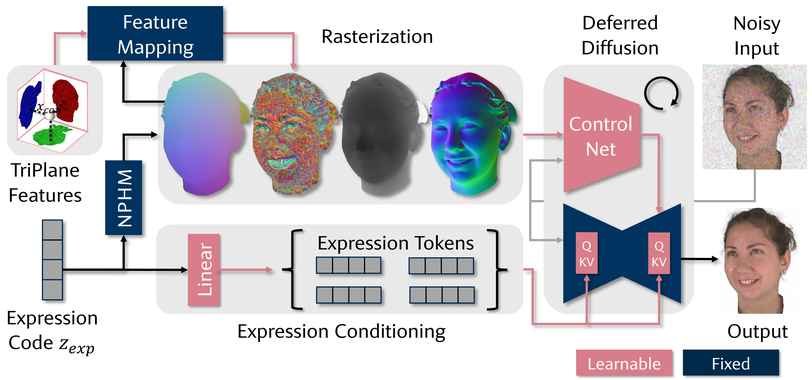
Input: Multi-view video recordings

Output: Animatable 3D Head Avatar

Motivation

Diffusion Models	3DMMs	**DiffusionAvatars**
✓ Photorealism	✗ Photorealism	✓ Photorealism
✗ 3D Control	✓ 3D Control	✓ 3D Control
✗ Expression Control	✓ Expression Control	✓ Expression Control

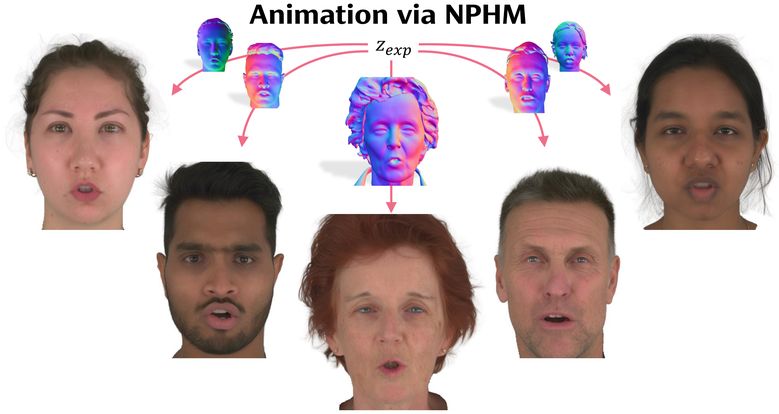
Method



2 streams to control expression:

- Screen-space decoding with ControlNet for 3D head pose + coarse expression
- Direct expression conditioning via cross attention for fine expression details

Animation via NPHM

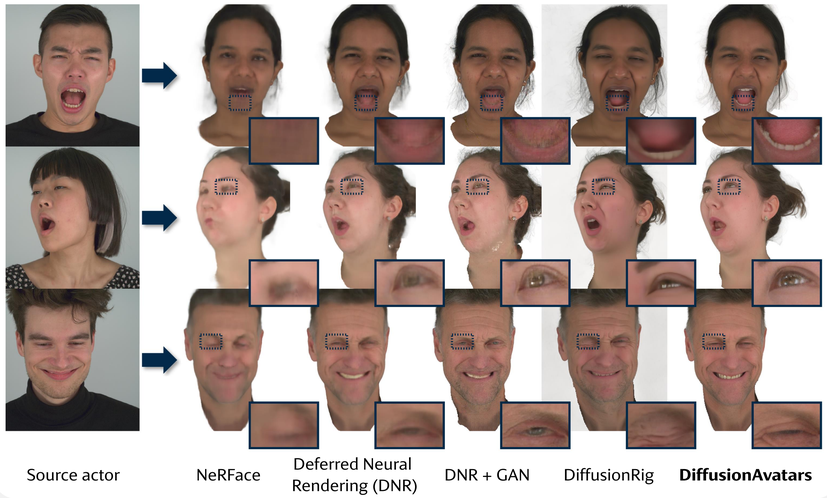


DiffusionAvatars can be animated by changing the input expression code z_{exp}

Expression Transfer

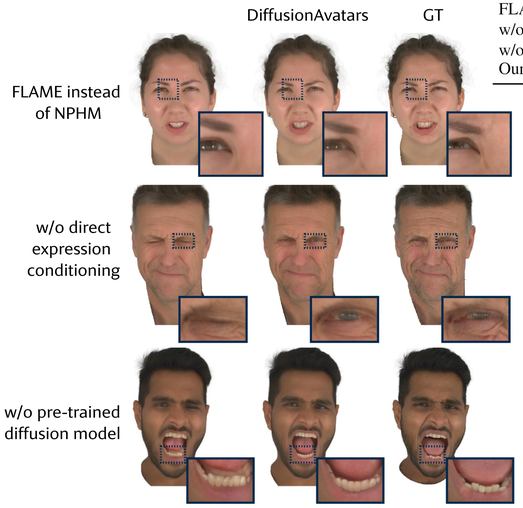
	NeRFace	DiffusionRig	DNR	DNR+GAN	Ours
Visual Quality↑	2.19	2.47	2.87	3.06	**4.02**
Driving Fidelity↑	2.35	3.52	3.94	3.97	**4.14**

User study



Ablation Study

	PSNR↑	LPIPS↓	JOD↑
FLAME	24.2	0.083	7.36
w/o exp. cond.	24.5	0.081	7.65
w/o LDM prior	24.5	0.078	7.67
Ours	**25.3**	**0.074**	**7.85**



NPHM instead of FLAME gives more detailed expressions

Direct expression conditioning helps distinguish expressions with similar 3D mesh

Using Pre-trained Diffusion Model steers generation towards more plausible images

Quantitative Comparison

Method	PSNR↑	LPIPS↓	JOD↑	AKD↓	AED↓	APD↓	CSIM↑
NeRFace	23.0	0.279	6.76	5.37	1.06	0.053	0.787
DiffusionRig	19.6	0.220	6.41	2.74	0.55	0.029	0.887
DNR	24.5	0.226	7.32	2.06	0.63	0.027	0.903
DNR+GAN	23.0	0.114	7.08	2.14	0.69	0.028	0.868
MVP	23.6	0.221	7.02	3.42	0.78	0.034	0.882
Ours	**24.9**	**0.081**	**7.55**	**1.79**	**0.50**	**0.023**	**0.917**

Self-reenactment results averaged over 8 persons

Project Page

Conclusion

DiffusionAvatars combines a detailed neural parametric head model (NPHM) with a diffusion architecture in a deferred neural rendering setting:

- Use ControlNet conditioned on NPHM renderings to steer 3D head pose + coarse expressions
- Exploit pre-trained diffusion model to produce more plausible images
- Use direct expression conditioning akin to IPAdapter for fine expression details
- Use Neural Textures stored in TriPlanes indexed via NPHM's canonical space for improved image synthesis