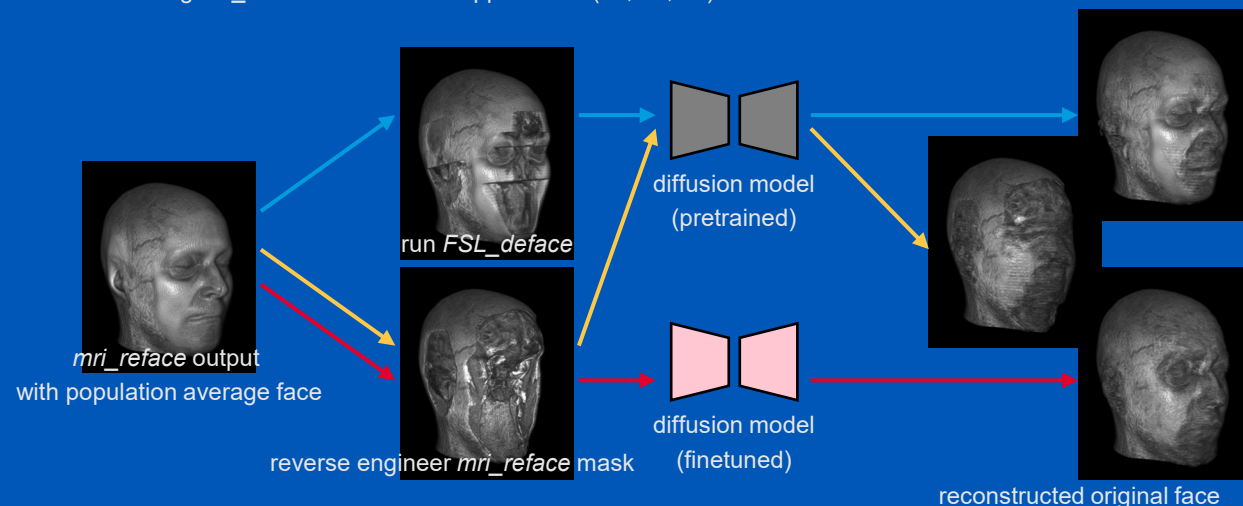




Replacing MRI facial voxels with realistic synthetic data (*mri_reface*) protects privacy against AI facial reconstruction attacks.



Reversing *mri_reface* with 3 attack approaches (#1, #2, #3)



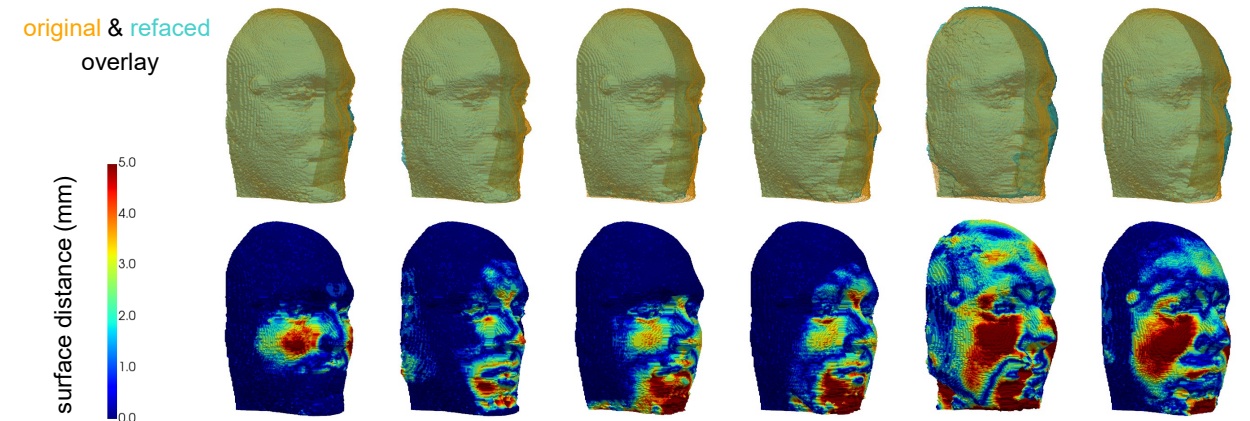
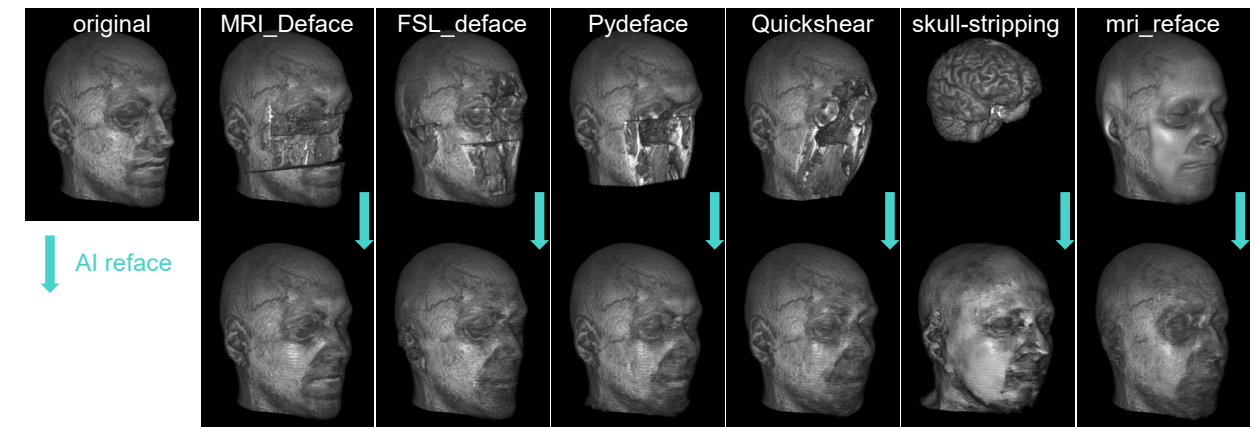
©2026 Mayo Foundation for Medical Education and Research

Red Teaming *mri_reface*: Assessing Reversibility via Diffusion-Based Facial Reconstruction Attacks

Chenyu Gao,¹ Michael E. Kim,¹ Yihao Liu,¹ Kaiwen Xu,¹ Trent M. Schwartz,¹ Lianrui Zuo,¹ Zhiyuan Li,¹ Derek B. Archer,² Timothy J. Hohman,² Ann Zenobia Moore,³ Luigi Ferrucci,³ Lori L. Beason-Held,³ Susan M. Resnick,³ Christos Davatzikos,⁴ Jerry L. Prince,⁵ Christopher G. Schwarz,⁶ Bennett A. Landman¹

¹Vanderbilt University; ²Vanderbilt University Medical Center; ³National Institute on Aging, National Institutes of Health;

⁴University of Pennsylvania; ⁵Johns Hopkins University; ⁶Mayo Clinic



BACKGROUND

- Facial structure in head MRI poses re-identification risk.
- Traditional defacing removes facial voxels; *mri_reface* replaces the face with a realistic, synthetic alternative.
- Generative AI can partially reverse some traditional defacing, but whether such reversibility extends to *mri_reface* is unknown.
- We conducted a red-team evaluation to test if *mri_reface* can withstand state-of-the-art AI reconstruction attacks.
- We tested 469 *mri-refaced* images against three escalating attack strategies, including reverse-engineering the mask and fine-tuning diffusion probabilistic models (DPMs, see workflow on left).

RESULTS & CONCLUSIONS

- While traditional defacing is easily reversed, all three reconstruction attacks on *mri_reface* failed.
- Pre-trained AI models were tricked by residual synthetic features (#1) and failed to generalize when aggressive cropping was applied (#2).
- Even using a fine-tuned model (#3), the faces were no closer to the original subject than a generic population average.
- Replacing facial voxels provides more robust privacy protection compared to naive voxel removal.

