PercHead: Perceptual Head Model for Single-Image 3D Head Reconstruction & Editing

Supplementary Material

A. Supplementary Video

We highly recommend watching our supplementary video, which showcases additional 3D reconstruction orbit views, frame-by-frame 3D video generation, 3D edit orbit sequences, and a live demo of our interactive 3D editing web application.

B. Perceptual Losses During Early Training

In Fig. 9, we compare early-epoch behavior (epochs 1–3) of different perceptual losses: LPIPS+L1, individual DINOv2 layers (2, 5, 8, 11), DINOv2 (8+11), SAM2.1, and our full combined loss. LPIPS+L1 struggles to reconstruct structures and high-frequency areas: glasses of subject 1 remain vague and incomplete even at epoch 3, teeth in subject 1 are not visible at all, and hair in subject 2 has poor texture across all three epochs. In contrast, DINOv2-based losses generally reconstruct glasses much earlier, often already at epoch 1 and more clearly thereafter. DINOv2 layer 2 is overly low-level, producing blurry outputs and without any glasses or hair structure; layers 5, 8, and 11 (and their 8+11 combination) capture glasses, teeth, and hair structure more reliably. However, these layers also introduce a red tint in the reconstructions. We also note that layer-5-supervised reconstructions are quite sharp, but we still opted for using later layers in our final loss formulation, as they offer more realistic hair. These layers are a good addition to the supervision of SAM2.1, which reconstructs details, including glasses and teeth, very well but is weaker on hair and introduces a yellowish tint. The combined DINOv2 (8+11) and SAM2.1 supervision yields: strong hair texture, robust glasses and teeth reconstruction, and without the tints observed when using either model alone.

In Fig. 10, we further analyze DINOv2 layer 8 supervision by comparing patch-only, class-only, and the combined patch+class setup after a single epoch. Patch-only supervision produces globally coherent geometry but also noticeably blurry areas (e.g., the chin of subject 2). Supervision based on the class token alone yields some sharper, stylized textures, consistent with our observations when training the SHP-CNN, but lacks geometric guidance: since the class token provides no spatially localized information, reconstructions become unstable, lose global 3D consistency, and exhibit color bleeding (e.g., red regions spreading across the face). As a result, class-only supervision is not enough to train the full 3D pipeline from scratch. Combining patch and class information provides localized geometric signals together with sharper textures, and is therefore used in our final configuration.

C. Additional Reconstruction Results

C.1. Comparison with FaceLift

FaceLift [47] is a single-image 3D face reconstruction method, which employs a diffusion-based multi-view generator trained on mutli-view synthetic head images. In Tab. 4 and Fig. 14, we compare our method against FaceLift. Quantitatively, our approach outperforms

Our approach outperforms FaceLift both quantitatively and qualitatively. In particular, our approach handles non-frontal inputs significantly better than FaceLift, which will often produce distorted reconstructions.

C.2. Quantitative NeRSemble Results

We report quantitative results on held-out NeRSemble [34] identities in Tab. 3 for both the novel-views and extreme-views tasks. For the extreme setting, we additionally consider strong vertical viewpoint changes, (i.e., bottom-input-to-top-target or vice-versa). The performance mirrors our observations on Ava-256: Except for SSIM, our method consistently outperforms baselines in perceptual, identity and pixel-wise metrics.

C.3. Qualitative Results on Ava-256 and NeRSemble

In Fig. 11, we present additional qualitative results on Ava-256 [48] and NeRSemble [34]. The first five samples correspond to Ava-256 subjects, and the last four to left-out NeRSemble identities. Across all cases and view transitions, our method reconstructs significantly more consistent heads than competing methods, with fewer artifacts around challenging regions such as hair, ears, and jawline, and with improved identity and shape preservation under large viewpoint changes.

C.4. Qualitative Comparison Between Target Views

Figures 12 and 13 analyze the behavior of each method when varying the target viewpoint for a fixed input image from Ava-256 [48].

In Fig. 12, we condition on a frontal view and render a sequence of target angles. Most baselines provide strong frontal reconstructions but quickly deteriorate when rotating away from the input view, revealing limitations in their underlying 3D consistency. Our method maintains stable identity and geometry across the full range of target angles.

Method	**NeRSemble** (Novel Views)					**NeRSemble** (Extreme Views)				
	PSNR ↑	SSIM ↑	LPIPS ↓	DS ↓	ArcFace ↓	PSNR ↑	SSIM ↑	LPIPS ↓	DS ↓	ArcFace ↓
LGM	11.4	0.662	0.500	0.189	0.701	9.9	0.641	0.563	0.243	0.758
PanoHead	15.4	0.750	0.261	0.106	0.338	14.4	0.716	0.310	0.119	0.373
SphereHead	15.3	0.758	0.260	0.106	0.336	14.3	0.720	0.321	0.135	0.433
GAGAvatar	16.9	**0.798**	0.217	0.107	0.288	14.6	**0.750**	0.317	0.162	0.476
LAM	14.5	0.758	0.303	0.134	0.353	11.6	0.682	0.433	0.204	0.551
Ours	**18.4**	0.764	**0.190**	**0.071**	**0.256**	**17.7**	0.741	**0.221**	**0.081**	**0.257**

Table 3. **Novel and Extreme View Reconstruction Performance on NeRSemble (Unseen Identities) [34].**

Ava-256 (Novel Views)	PSNR↑	SSIM↑	LPIPS↓	DSim↓	ArcFace↓
FaceLift	12.5	0.645	0.436	0.129	0.547
Ours	**16.4**	**0.691**	**0.269**	**0.092**	**0.292**
Ava-256 (Extreme Views)	PSNR↑	SSIM↑	LPIPS↓	DSim↓	ArcFace↓
FaceLift	11.0	0.627	0.482	0.181	0.556
Ours	**15.9**	**0.678**	**0.291**	**0.106**	**0.282**

Table 4. **Comparison against FaceLift [47] on Ava-256 [48].**

In Fig. 13, we use a side-view input. Here, some baselines already struggle on the input view and introduce noticeable artifacts when extrapolating to frontal and opposite-side views. Our method remains robust for both the input and unseen angles.

C.5. Robustness Analysis

In-The-Wild Scenarios and Generated Inputs

Figure 15 investigates robustness to in-the-wild and generated images. We evaluate our model on left-out FFHQ [31] samples and diffusion-generated images with cartoon-like, stylized, or otherwise complex appearances. Despite substantial domain shifts in pose, lighting, and style, our method continues to produce plausible 3D heads and preserves identity. These results suggest that the learned representation generalizes beyond the training distribution and remains applicable in realistic and creative downstream scenarios.

In Fig. 16, we further demonstrate that our method can produce complete side views and retains good performance even in difficult out-of-distribution scenarios with strong shadows and difficult illumination. However, we do observe that difficult geometry, such as exotic glasses, are not perfectly reconstructed by our model.

SHP-CNN in Out-Of-Distribution Scenarios

In Fig. 17, we that our SHP-CNN, which was only trained on Nersemble [34] and VFHQ [74] samples, also handles synthetic Cafca [6] samples well. We also show an in-the-wild image with difficult lighting. Our SHP-CNN improves perceptual fidelity primarily by adding sharpness and stronger contrast. However, it does not add any new details. These are generated solely by the reconstruction model.

Loss Robustness

We compare our loss formulation and LPIPS+L1 in Fig. 18, with deactivated SHP-CNN. Our loss shows strong robustness to difficult out-of-distribution inputs and viewing angles, while the standard loss leads to collapsed reconstructions.

C.6. Computational Costs and Lightweight Model

In Tab. 5, we compare the performance and computational requirements of our model and two smaller variants: (a) one model with four layers and no SHP-CNN, but using the same DINOv2 (Giant) encoder as our standard model, and (b) a model with the same configuration as (a), but using the smallest DINOv2 encoder. Both models provide strong performance and offer lightweight alternatives to our base model.

Ava-256 (Extreme Views)	SSIM↑	LPIPS↓	DreamSim↓	ArcFace↓
Light-S (4 layers, no SHP, DINOv2-S)	15.4	0.313	0.120	0.316
Light (4 layers, no SHP, DINOv2-G)	**16.1**	0.298	0.110	0.285
Ours (10 layers, SHP, DINOv2-G)	15.9	**0.291**	**0.106**	**0.282**

Model	**DINOv2**	**Forward**	**GS**	**SHP**	**DINOv2**	**3D-Model**	**SHP**
Light-S	0.01 s	0.03 s	0.03 s	-	0.5 GB	2.1 GB	-
Light	0.02 s	0.03 s	0.03 s	-	5.3 GB	2.1 GB	-
Ours	0.02 s	0.07 s	0.03 s	0.04 s	5.3 GB	2.6 GB	6.1 GB

Table 5. **Performance, computational cost, and memory requirements of our lightweight and standard model.**

D. Disentangled 3D Editing

D.1. Additional Image- and Prompt-Based Editing Comparisons

In Figs. 19 and 20, we present additional qualitative results for disentangled 3D editing using both image- and prompt-based stylization.

D.2. 3D Editing Web Application

Our 3D editing web application exposes the full disentangled editing pipeline in an accessible interface. Users first upload an input image, from which the pipeline extracts a semantic segmentation map. The segmentation can then be interactively modified via drawing tools to, for example, reshape hair regions or add glasses.

Extracting a segmentation map from an image takes approximately 20 seconds, as it involves both GAGAvatar's preprocessing [9] and FARL [79] for semantic segmentation. This step is only required when uploading a new image; subsequent edits reuse the existing segmentation.

For stylization, users can either upload a reference image or provide a text prompt. Stylization with a reference image takes around 20 seconds, including CLIP-based feature extraction [58] and a forward pass through our model. In contrast, generation using a text prompt is significantly faster, requiring only about 10 seconds. Overall, the web application demonstrates that our disentangled 3D editing framework is not only flexible but also practical for interactive use.

E. Training & Architecture Details

Training. All models are trained on a single NVIDIA RTX 3090 GPU. The 3D reconstruction network is optimized for 15 epochs, corresponding to roughly 72 hours of training. Afterwards, the SHP-CNN head is trained for 30 epochs (about 10 hours) with the rest of the pipeline frozen, using only real data (i.e., excluding Cafca [6]). The disentangled editing model reuses some of the pretrained 3D reconstruction weights and is further trained for 24 epochs (about 48 hours) without Cafca to avoid bias towards generating artificial heads. We employ the AdamW optimizer with a learning rate of 3×10^{-5} and a weight decay of 10^{-5}.

Additional Architecture Details. We use a frozen DINOv2 ViT-g/14 encoder as a pre-processing stage, extracting four feature maps of dimensionality 1536 from layers 9, 19, 29, and 39 (similar to LAM [28]). The second encoder branch is a lightweight 2-layer ViT encoder with a dimensionality of 128. The decoder is a 10-layer ViT with a dimension of 1024 and operates on 4096 patches, each representing 16 Gaussians. We employ 16 cross-attention heads per decoder layer.

F. Evaluation and Processing Details

F.1. Subjects used for quantitative evaluation:

NeRSemble [34]: 059, 070, 370, 373, 374
Ava-256 [48]:
- `20220809--1034--BJM420`
- `20220815--1307--BMP511`
- `20220831--0751--CMS162`
- `20230224--1359--CMZ386`
- `20230308--1352--BDF920`
- `20230316--1103--BHK376`
- `20230324--0820--AEY864`
- `20230328--0800--BLY735`
- `20230405--1635--AAN112`
- `20230810--1630--ANX726`
- `20230914--1105--BXQ083`

Cropping Alignment **PanoHead** [1] uses the tightest (smallest) image crops among all compared methods. To ensure fair and consistent evaluation across models, we applied the same PanoHead cropping to all methods for qualitative and quantitative comparison.

GAGAvatar Processing For **GAGAvatar** [9], their default rendering pipeline produces images with a black background. To standardize appearance and ensure comparability across methods, we replace the black background with white using their official `GAGAvatar Track` [9] preprocessing pipeline.

G. Note: Visualization of Intermediate Decoding States

In our method overview (Fig. 2), we visualize intermediate reconstruction states in our ViT decoder. For each visualization, we run a full forward pass, but control the activation of the cross-attention mechanisms. Specifically, to visualize the output after decoder layer i, we keep all cross-attention layers active up to and including layer i, while disabling cross-attention for all subsequent layers. Importantly, we retain the MLP blocks and skip connections in all layers, ensuring that feature propagation and refinement still occur. This setup allows us to isolate the contribution of 2D feature retrieval up to a specific depth in the decoder.

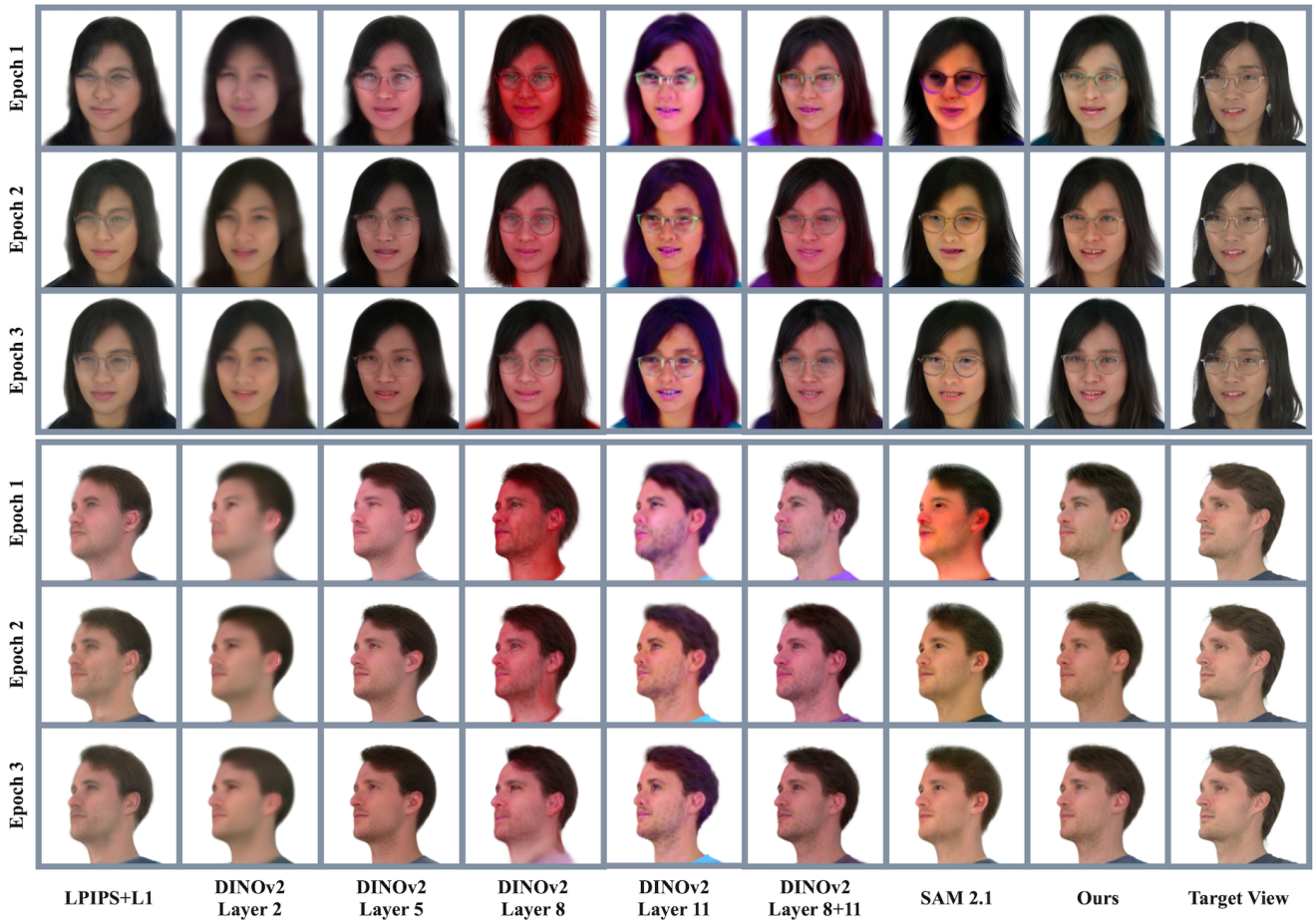

Figure 9. **Early-Epoch Comparison of Perceptual Losses.** We visualize epochs 1–3 for LPIPS+L1, different DINOv2 layers, SAM2.1, and our full model. DINOv2+SAM2.1 provides the best supervision, recovering glasses, hair, and teeth structures much earlier and with fewer color artifacts than the alternatives.

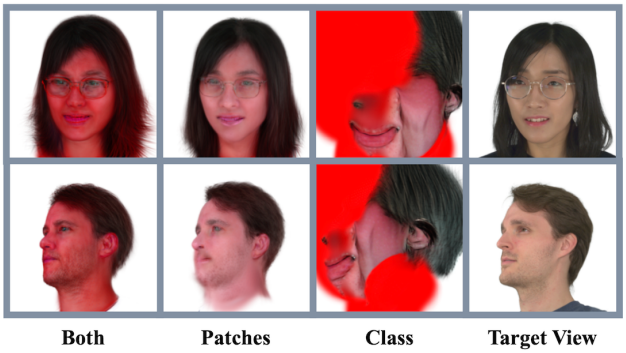

Figure 10. **Patch vs. Class Token Supervision (DINOv2 Layer 8).** After one epoch, patch-only supervision yields coherent but blurrier geometry, while class-only supervision produces sharper yet geometrically unstable reconstructions. Combining patch and class information offers best guidance.

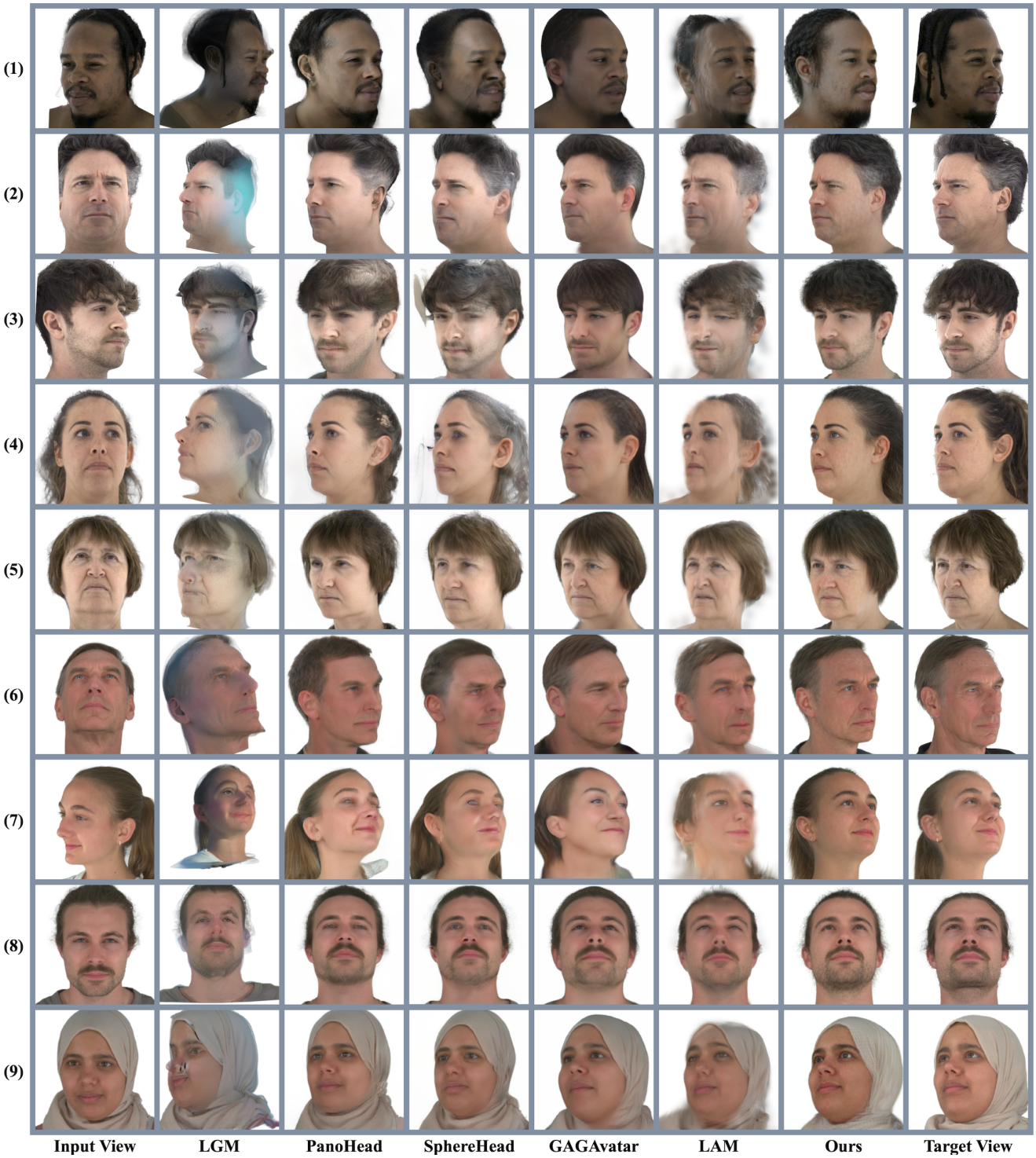

Figure 11. **Additional Qualitative Results on Ava-256 [48] and NeRSemble [34].** We show additional cross-view reconstructions on held-out subjects. Samples (1)–(5) are from Ava-256, while samples (6)–(9) are from left-out NeRSemble identities.

Figure 12. **Multi-View Reconstruction from a Frontal Input on Two Samples from Ava-256 [48].** Starting from a single frontal input view, we reconstruct a wide range of target viewpoints on a NeRSemble subject. While most methods typically perform well on the frontal view but degrade under large view changes, our method maintains consistent identity, facial geometry, and hair structure across all target angles. This demonstrates that our model genuinely recovers a stable 3D representation rather than overfitting to the input view.

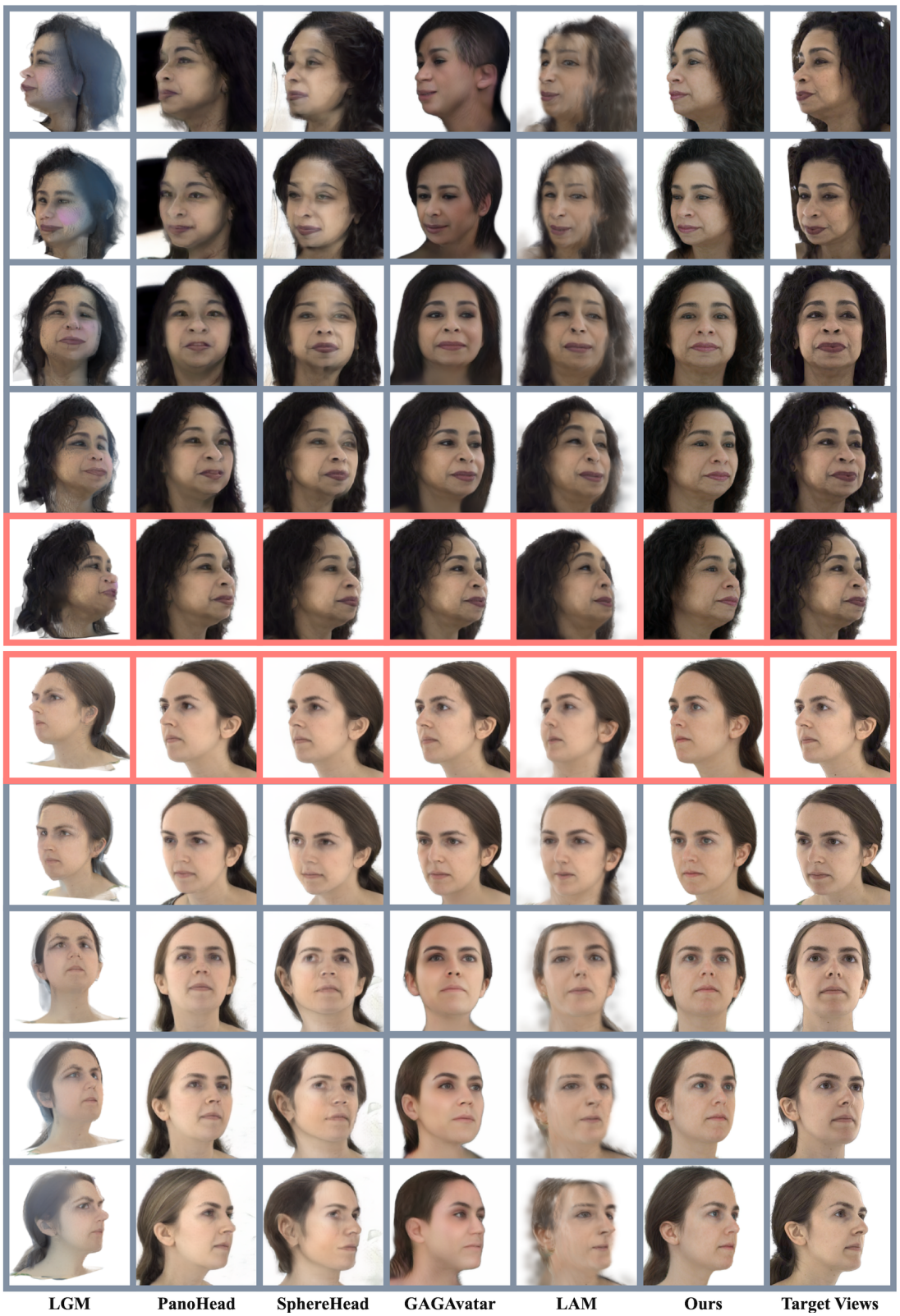

Figure 13. **Multi-View Reconstruction from a Side-View Input on Two Samples from Ava-256 [48].** We repeat the multi-angle reconstruction experiment using a challenging side-view input. Baseline methods often struggle to generate plausible frontal and opposite-side views, leading to identity changes, distorted shapes, or collapsed geometry. In contrast, our method produces coherent reconstructions for the input side view as well as unseen target angles, indicating strong 3D consistency and robustness in self-occluded facial regions.

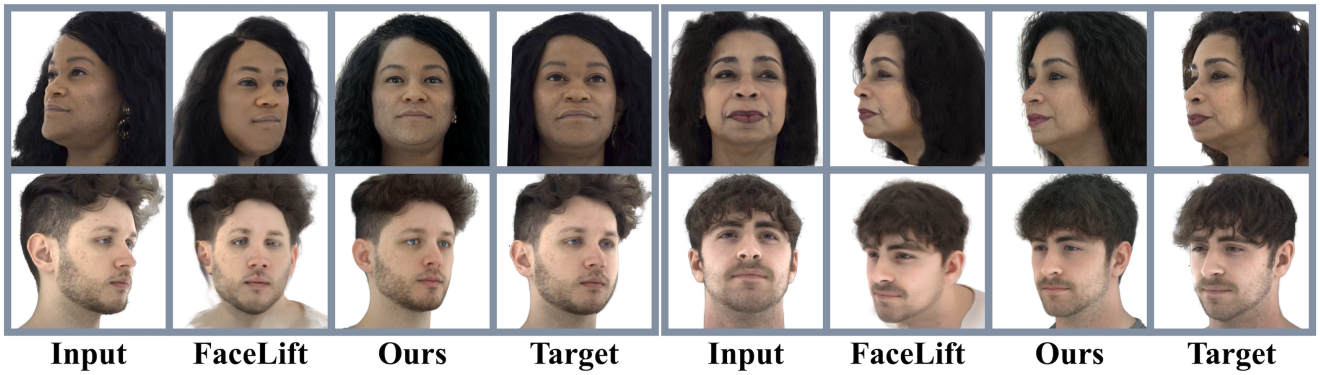

Figure 14. **Qualitative comparison with FaceLift [47] on Ava-256 [48].**

Figure 15. **Robustness on In-the-Wild and Generated Images.** We evaluate robustness on challenging inputs beyond the training distribution. The first block shows left-out FFHQ [31] samples, while the second block shows reconstructions of images synthesized by a diffusion model (SORA [50]), including cartoon-style, highly stylized, or otherwise complex appearances. Across both categories, our method offers extremely robust performance. Due to the training data, it does however lean towards more realistic reconstructions, even with cartoonish inputs. We also see that heavily unrealistic faces can sometimes reduce reconstruction performance - like the right sample on the second-to-last row.

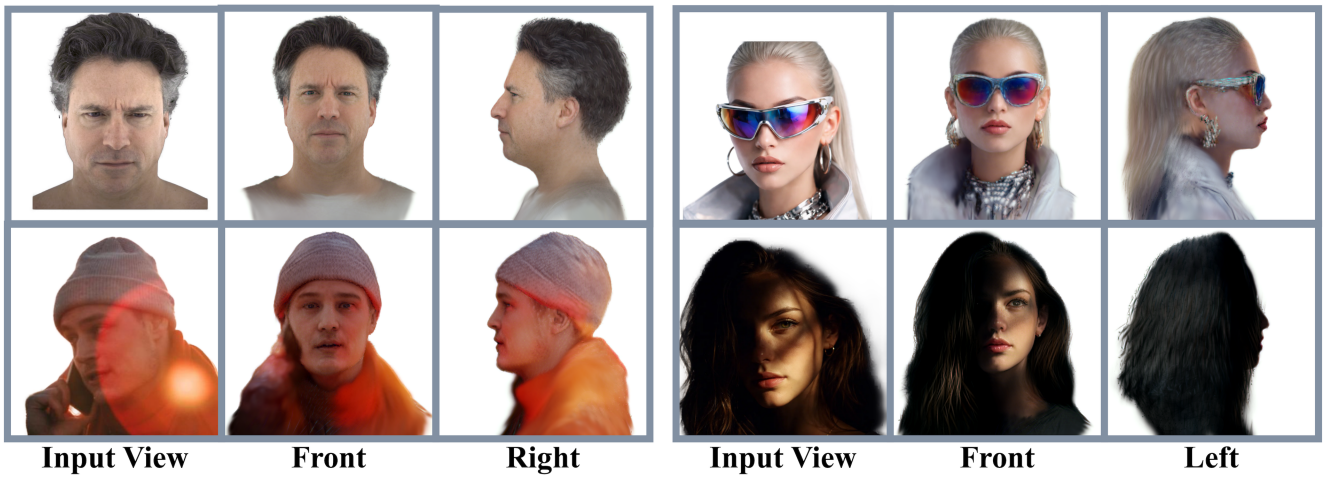

Figure 16. **Extreme Scenarios and Side Views.** Inputs are from Ava256 [48], SORA [50], and UFDD [49].

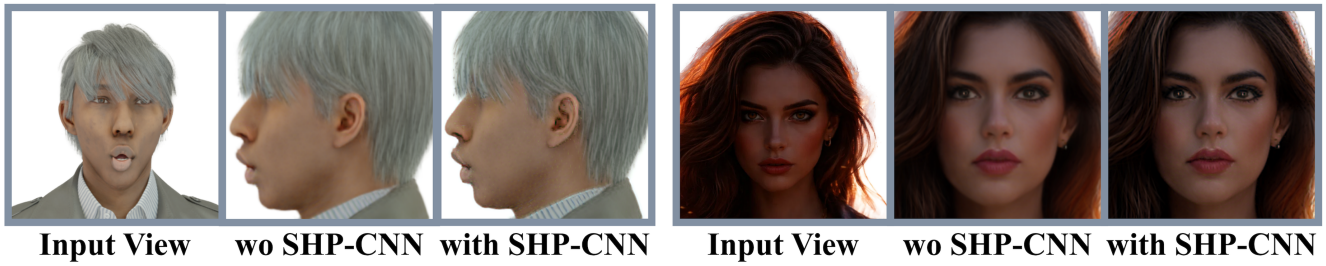

Figure 17. **SHP-CNN effects for out-of-distribution sample and difficult lighting.** Inputs are from Cafca [6] and SORA [50].

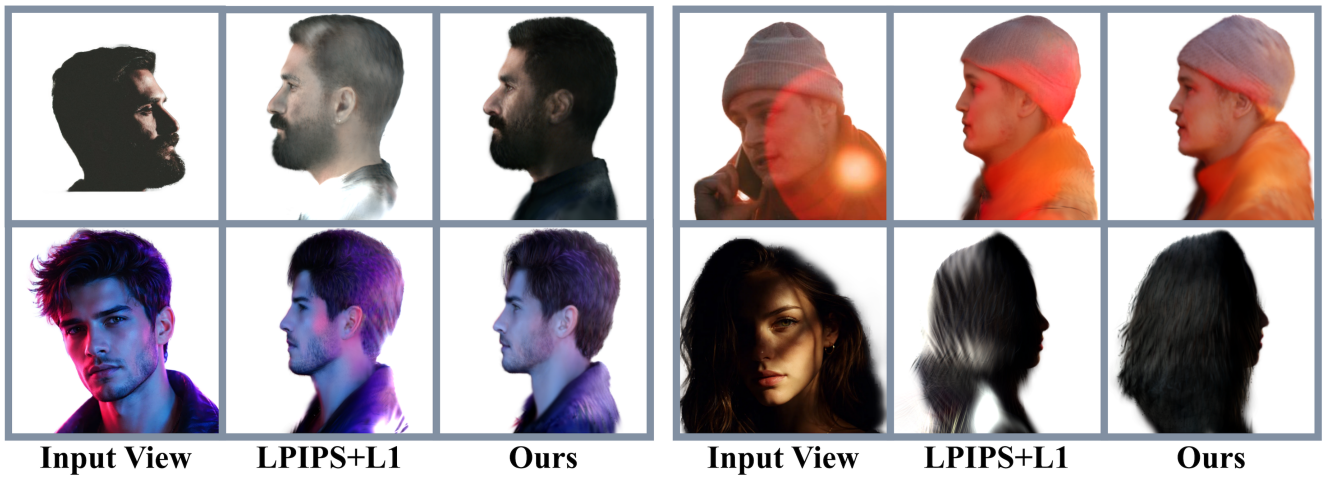

Figure 18. **Robustness of Our Model and an LPIPS+L1-Trained Ablation, Both Without the SHP-CNN.** Inputs are from UFDD [49] and SORA [50].

Figure 19. **Additional 3D Editing Results with Style Guided by Image.**

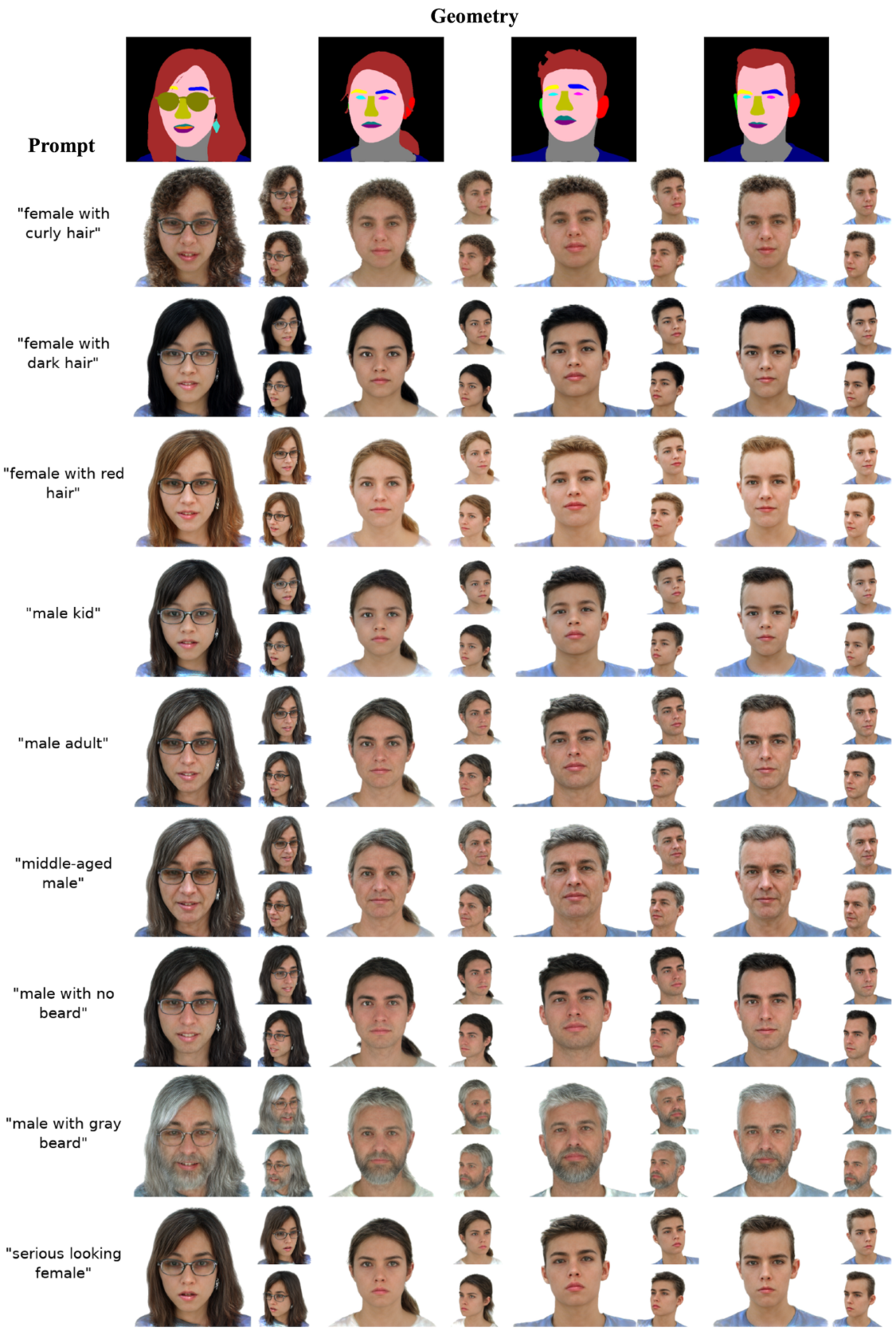

Figure 20. **Additional 3D Editing Results with Style Guided by Prompts.**